GREEN UNIVERSITY OF BANGLADESH (GUB)

# Hybrid Approach to Automated Contextual Story Generation in Bangla


**Submitted by**

Md. Fahad Hossion (201002048)

Likhan Mia (201002340)

Saidur Rahman (201002142)


*A thesis submitted to the Department of Computer Science & Engineering*
*for the partial fulfillment of the degree of*
*Bachelor of Science in Computer Science & Engineering*


**Supervised by**

Mr. Tamim Al Mahmud

Assistant Professor, Department of Computer Science and Engineering



Department of Computer Science & Engineering

Green University of Bangladesh

Purbachal American City, Kanchon 1460



February, 2024

# Declaration

We declare that the submitted thesis is entirely based on our own research. Every work here is authentic and references are provided to previous works from which we were inspired. This thesis, neither in whole nor in part, has been previously submitted for the award of any degree.

| | | |
|---|---|---|
| Md. Fahad Hossion | Likhan Mia | Saidur Rahman |
| ID: 201002048 | ID: 201002340 | ID: 201002142 |

# Certificate

This is to certify that the thesis entitled **Hybrid Approach to Automated Contextual Story Generation in Bangla** has been prepared and submitted by **Md. Fahad Hossion, Likhan Mia , and Saidur Rahman** in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering in February 2024.

---

Mr. Tamim Al Mahmud
Supervisor

Accepted and approved in partial fulfillment of the requirement for the degree Bachelor of Science in Computer Science and Engineering.

---

Rokeya Khatun
Member

---

Wahia Tasnim
Member

---

Jahidul Arafat
Member

# Acknowledgments

We firstly express our gratitude and thanks to the most merciful almighty Allah who owns all powers and by whose will we have been able to do this work and learn something new.

Then we express our thanks to our thesis supervisor Mr. Tamim Al Mahmud Sir, Assistant Professor, CSE Department, Green University of Bangladesh. His cooperation made our work easier whenever we sought his help in every need. Despite his busy schedule, he spent his precious time behind us and we have been able to complete our thesis.

Thereafter, we would like to thank our family, friends, and all the well-wishers who have always stood by us and inspired us. Who gave us a clean and beautiful life. The family who have supported us financially as well as also gave us mental strength and motivation to move forward and do our best.

Next, we would like to thank the respected faculties of our Defense Board, as they have given us more understanding by correcting mistakes, and through their judgments we have been able to motivate us to do better.

Finally, we would like to thank our honorable chairman of the CSE Department, Associate Professor Dr. Muhammad Aminur Rahman Sir, and our institution Green University of Bangladesh, for allowing us to do this research work and to complete our B.Sc. degree in Computer Science and Engineering.

# Abstract

Story generation systems are a form of text generation technique in computer science. In this system, a corpus of stories is needed which essentially helps to generate the text and generates it randomly or based on user input. Natural Language Understanding (NLU) techniques which are part of NLP must be applied to understand user input. In our research we mainly focus on Natural Language Generation (NLG). As a result, the generation process gets more time to generate correctly. The Bengali story is an unstructured one from the point of view of AI technology due to its lack of research, which we faced most of the problems in our research. We have proposed new approach for text generation specifically story create although these are not sufficient works for Bengali story generation. We call it bridNG which is a combination of neural network model RNN(GRU) and traditional algorithm Ngram and it works based on hidden markov model technique. Here we use HMM, Markov Model, Gated Recurrent Unit and pre-trained model to generate stories and trained them with our own dataset collected where there is no previous Bengali story dataset. After generating stories using all models and algorithms, the proposed model gives more coherent sentences (gives a score of 52.23% by human evaluation) than others by evaluating human and pre-trained models.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Story Generation is a creative system of text generation in **NLP** (Natural Language Processing) that helps create coherent and fluent passages of text about a topic. In the 1960s and 1970s, researchers began experimenting with simple text generation programs, and a notable example is the program **Racter**, developed in the early 1980s, which could generate poetic and often surreal text. After 10 years, expert systems became popular in artificial intelligence (AI) research where systems use rules and knowledge bases to mimic the skills of human experts in various domains. In recent times it refers to the process of using AI or computational algorithms to autonomously create narratives or stories. It began in the late 20th and early 21st centuries, as machine learning gained prominence, researchers discovered statistical and probabilistic techniques for text generation, including **RNNs** expert neural networks. Also among these methods, the Markov model and later the transformer model, gained popularity for their effectiveness in text generation. In modern era, With the rise of deep learning and powerful computing resources, large language models **(LLMs)** have achieved a dramatic leap in story generation capabilities.

The generation of story is important in several works, with most creativity & exploration being one of them. In this sense it allows the exploration of new narratives & ideas and enables writers, creators and even AI systems to create

diverse and innovative stories. We are working on the Bengali language to generate more efficient and coherent story. This language spoken by over 250 million people worldwide as their first or second language[1]. But there has been very little research on Bengali language. We are working on the generation of stories in this language to enrich the research of Bengali language. Although this language is easy and understandable for us who speak this language, it is not at all easy for machines or computers. The controllability issue in story generation is the user's ability to influence the outcome of input generation. Such influence often takes the form of a plot the user wants the system to adhere to while creating a new narrative[2]. This is still an open challenge for automatic story generation in the text generation task. So, our big challenge is to make the machine understand the plot of stories.

We collect Bengali stories and analyze the generation process where it differs from stories of other cultures and present here our social and cultural activities that we can easily recognize. The history of Bengali stories is a vibrant tapestry woven by ancient myths, religious epics, folklore, colonial influences and contemporary expressions[3]. For these different challenges of Bengali story generation, we used both Machine Learning (**ML**) and Deep Learning (**DL**) algorithms to analyze the results and compare their generation output. It is notable we collect stories to build our dataset from different sources and here problem is collected stories has various thought from the writers. The training process of each model of Bengali story generation is more complex than other models and the analysis process takes more time, but the generation process is simpler than other tasks.

## 1.1 Motivations

The research on automatic Bangla story generation is driven by a number of factors, including the necessity for cultural preservation and technological progress.

**i. Bengali culture preservation and promotion:**

Bengali literature and cultural legacy are abundant. But there's a chance of cultural dilution given the prevalence of English and other languages in the digital and technological spheres. Researchers hope to use technology to conserve, promote, and reinvigorate Bengali literature, folklore, and cultural narratives by concentrating on automatic Bangla story generation. The younger generation and Bengalis residing around the world may now access Bengali culture thanks to this initiative, which also ensures the language's continuous progress in the digital age.

**ii. Improvement of Natural Language Processing (NLP) in Bengali:**

While several languages have witnessed notable progress in NLP, Bengali has gotten relatively less attention. Researchers hope to bridge this gap by creating specialized NLP algorithms for Bengali, especially in story development. This entails learning the subtleties of the language, including its syntax, semantics, and colloquial idioms, and then applying that knowledge to algorithms that may produce compelling narratives. Improving NLP skills for Bengali provides real-world uses in education, entertainment, and information sharing in addition to serving scholarly and technological goals.

**iii. Promoting Research and Innovation in Bengali Language Technologies:**

Although language technologies are of global importance, Bengali has witnessed little innovation in this field, despite being one of the most spoken languages in the world. The goal of the research is to increase interest in and funding for Bengali language technology by concentrating on automatic story generation. This could stimulate innovation, create a more dynamic ecosystem of tools, applications, and services for Bengali speakers, and motivate more scholars and developers to work on initiatives involving the Bengali language.

**iv. Creating Priceless Resources for Education and Learning:**
Automated story generation can generate a multitude of resources that educators can employ to enhance students' learning experiences. Narratives are an effective tool for imparting moral lessons, presenting cultural contexts, and teaching language. Teachers can gain access to a wealth of material that can be utilized to teach reading, writing, and comprehension skills in a way that is engaging and culturally appropriate by compiling a broad dataset of Bengali stories.

## 1.2 Aims and Objectives

In our research, we aim to enrich the Bengali texts by developing special natural language processing techniques tailored for Bengali. Story generation for Bengali is still under-researched in Bangladesh although other technology companies are re-researching it through other languages. We work with story generation from Bengali to improve the Bengali language model to get some unique insights into reader stories.
Our main objectives are:

- To generate an own Bengali stories dataset that helps other research.
- To improve intelligence Tasks for the Bengali Language.
- To create different types of stories via multiple models and get more effective results.
- To increase the number of research in the Bengali language.
- To develop an effective model that can be used for generational work for Bengali text.

## 1.3 Research Questions

Through our work, we will answer the following research questions:

- How does the given generation technique generate a story?
- How are the benefits of this research helpful in real life?
- Can we propose a new technique for story generation for Bangla?

## 1.4 Research Contribution

For story creation, a dataset is created by collecting data from different sources and creating stories through different models. We took the help of various people to evaluate the results.

Stories made in Bengali text can help Bengali literature lovers who will get some new context to read. Also it will contribute to our culturally engaging content creators, game plot creators, filmmakers and etc. But here all the stories have to be clarified as modern literary standards. Here we have worked out the new generation technique how it produces more coherent stories and the proposed technique gives better generation output than other models used in our study. It is a hybrid approach of RNN and Ngram algorithm and they are deep learning algorithm and machine learning algorithm. In this technique it always tries to produce the best output based on the trained data corpus. It produces the best because it always compares both predicted next words of each model and we get more coherent sentences. In our method evaluation phase, we used human evaluation scores on the generated stories to get approximate evaluation scores where more than 20 people help us get the scores. In the final stage of research, the results of our model are evaluated by three main categories and conditions, sentence structure or grammar check, story coherence as well as context or prompt or story theme.

## 1.5 Thesis Outline

- **Chapter 2**: Literature Review in various languages describe significant previous research in the field of text generartion and story generation.
- **Chapter 3**: Methodology part which describes in detail our whole work and it also represents our proposed algorithm with the system architecture.
- **Chapter 4**: Performance Evolution is Where the performances of our applied machine learning and deep learning models are shown.
- **Chapter 5**: The conclusion shows the overall performance of the system, limitation and prime difficulties faced during working.

# Chapter 2

# Literature Review

## 2.1 Introduction

Research on story generation outside the English language is not promising. There are some studies in other languages such as Chinese that are available for publication on the Internet but most of the discussions are not documented.

Story generation is a type of text generation technique and we collect various text generation techniques in different languages such as Bengali, English, Chinese or Mandarin, French and previous works in other languages. We used these papers to enrich our research models to generate stories in the best possible way. Here is a discussion of the research paper to find a correlation with our research.

## 2.2 Work In Bangla

Bengali story generation-related papers are not promising for a few research on the Bengali language. In the paper **Sequence-to-sequence** Bengali Sentence Generation, they used Long Short Term Memory (LSTM) a special type of RNN architecture[4]. The authors of the paper used neural networks to train with a Bengali newspaper corpus in which they collected 917 days of newspaper text from the online Bengali news portal Prothom Alo. Here RNN neural network attempts

to model sequence or time dependent in regular behavior and the author of the paper follows some approaches where it is done by output feeding back of a neural net layer in time t to the input layer at time. They use Long Short Term Memory (LSTM) because it is capable of learning long-term dependency problems. The neural network is made to avoid the long-term dependency issue and keeping information for long periods is their actual default behavior, nothing that they struggle to be trained.

Working with Bengali is still very difficult and the processing stage of Bengali text data is a difficult task as they are very noisy and not suitable for working with machine learning or deep learning methods. For this reason, they did some preprocessing work to denoise our dataset and make it perform optimally in the neural networks, such as removing all Bengali punctuation marks, extra spaces, and new lines. After this process, they converted the text to utf-8 format[4]. Next, they trained their model for only one week's news paper corpus for having their limitation of the hardware and finally, they did a test with different Bengali text, then their model generated some text according to their given previous text.

Another paper on Bengali text generation where they use Bidirectional RNN[5]. They talked about N-gram modeling with pre-trained Bengali word embedding for text generation and after that, they built a bi-directional recurrent neural network to prepare their model. In their dataset collection stage, they used their dataset which was collected from online life where the dataset contains several types of Bengali posts such as collection posts, individual posts, page posts, and more. For Snag Bangla data collection, they try to reduce the majority of those obstacles to keep pure Bengali content in their dataset. Their datasets contain textual information and their arrangement and content outlines or summaries present in their datasets. After they use pre-trained word embedding for Bengali for the requirement their model is to convert normal language to vector. Here they use **bn_w2v_model** for their research purpose which gives them more accurate output from other presented pre-trained word embedder for Bengali. In the

next process, they work with the tokenized model which concentrates words with their record number from the corpus of their dataset, and all content changes the arrangement of the token. Later, for alternate lengths of each progression, they used the pad sequence and Keras pad sequence function to scale the length of the arrangement. In the promise learning model they used n-gram gathering as their given word and predicted word as associated word.Their proposed model RNN has two directions one forward and one backward, both opposite directions. The model gives information from forward and backward on the output density. Here the past or previous information refers to the backward direction and the subsequent or predictive sequence refers to the forward direction. In their model they used two activation functions such as ReLu and softmax where Rectified Linear Unit is used to activate the LSTM cell in bi-directional RNN. As a result, their trained model gave it an accuracy of 98.766% to predict next word in about 3 hours and their model loss was 0.0430.

The conditional language model is used in our last studied paper of Bengali text generation[6]. In the last paper, they proposed a Bidirectional gated recurrent unit (GRU) based architecture where it simulates the conditional language model or the decoder portion of the sequence to sequence (seq2seq) model. It is further conditioned upon the target context vectors of given data. Their proposed model had an evaluation metric based on human scoring and they used it to compare the performance of the model with unidirectional LSTM and GRU networks which are variants of RNN. They present a rhetorical overview of a successful work on context-driven text generation for Bengali language on which no established work of research has been conducted. They examined and analyzed several related topics, including context vectors, NN architectures, and inspections of optimization algorithms. Here they wanted to find the most effective and suitable combination. Final of their work, they propose a novel evaluation method that involves real-world feedback. Their novel evaluation method cannot be evaluated with existing methods. According to their proposed procedure, on average each generated se-

quence contained 70.86% of the expected features.

The framework proposed in the last paper was based on a generative conditional language model similar to the decoder of the autoencoder architecture[6]. In their study, given some input sequences from the vocabulary, their model uses an LM to transform these word representations into the vector space $p(x_0, x1, ..., x_{t1})$ and predict a probability distribution over the set V. Conditioned by some context words to indicate the overall meaning of the output sequence. This context is also provided as input to the model and transformed into a constant dimensional vector ct using the same LM. The output of the model is $p(x_t|c_t, x_0, x1, ..., x_{t1})$. They trained their model with two inputs—an embedding vector representing context and n-gram training sequences from the dataset. They also trained their model and optimized it first with 100 epochs on the full training data. To make the model more robust to a larger set of contexts, they generated new context vectors every 10 epochs.

## 2.3 Work In English

The knowledge of story generators is made up of several, intricately linked parts. An understanding of the information required by computational tale generators was proposed by Alhussain, Arwa and Azmi, Aqi [7]. A comparable but more expressive knowledge representation approach was put out by Alhussain, Arwa and Azmi, Aqi. They have used various types of models like Structural Model, Graph Based Approach, Grammar Based Approach, Planning Based Model, Goal Directed Approach, Analogy Based Approach, Heuristic Search Approach and Machine Learning Model. At first They are abstracting the story. Textual stories often include a lot of information that are unimportant to the storyline and provide the learning process additional dimensions. Consequently, it was imperative to develop a narrative abstraction that both enhances the likelihood of tale overlap and simplifies story representation. This made tales less sparse and promoted more effective learning and inference. The most popular application of tale ab-

straction was to simplify the story's events and highlight its primary characters and plot points. Then they have applied script learning of corpora. Using story corpora to generate tales began with script learning and creation. This approach seeks to ascertain the degree of relationship between a certain event and a group of occurrences. A statistical model like this can be used to forecast future occurrences that fit within a specific sequence of events. Then they have checked the story completion. The Children's Book Test (CBT) datasets was utilized as a tale corpus by Alhussain, Arwa and Azmi, Aqil. The process of creating a tale begins with an initial 20-sentence story. The following sentence in the story is then generated using CBR. Next, using the original 21st sentence as a gold standard, RNN is utilized to construct the final phrase word-by-word.

The primary contribution of the work of the research on English is the automated evaluation of the created stories through the use of many linguistic metrics. Furthermore, a context-aware hierarchical LSTM model that can forecast future sub events based on past sub events was proposed by Alhussain, Arwa and Azmi, Aqi. This model produces a list of phrases that will describe the upcoming sub event. The temporal sequence of events and the word sequence are the two layers of the event sequence that are taken into account. The tale subject is taken into account as an extra contextual aspect. Then they have started story generation. The effectiveness of Seq2Seq models in various NLP tasks encouraged researchers to employ them to create whole tales. Two commercial systems were integrated by Alhussain, Arwa and Azmi, Aqi. To create a story generator that, when given a series of separate, brief descriptions, creates stories. First, individual phrases within a sentence were translated using statistical machine translation (**SMT**). Subsequently, a deep RNN was employed to encode every sentence individually and subsequently decode them into coherent narratives. Despite the fact that this system could produce summaries that resembled stories, they lacked a complete semantic relationship to the input description. The assessment measures that were used yielded rather low overall ratings. At last they have done the story evaluation.

It is crucial for both controlling the generating process and assessing the created tale. Even though story creation systems have advanced significantly and are now capable of producing workable outcomes, narrative assessment lags behind and is still seen as a persistent issue. When compared to previous AI models, story creation expands the range of potential stories by adding subjectivity, diversity in evaluation criteria, and high dimensionality of narrative components. The authors have used average rank and Recall@N formula to evaluate the story. Decomposition, Deep learning, Hybrid systems, Automatic evaluation, Bench marking are area to upgrade for the paper published by Alhussain, Arwa and Azmi, Aqil and their accuracy parentage is 64%. The issue of automatic story generation has been approached in a variety of ways. These vary in the data sources and techniques they utilize to create their narratives. They provide an overview of these strategies and the systems that have been built up to this point in the first half of this chapter. These methods of creating stories fall under several categories, including problem-solving, agent-based systems, narrative grammars, story schema, and commonsense knowledge. Their major goal is to assess these systems' scalability, namely their ease of transfer ability from one domain to another. Since the goal of this thesis is to create textual tales, they must investigate the ways in which knowledge bases may be used to construct text documents. They will concentrate on NLG systems that have undergone data-driven training, hence eliminating the need for manually constructed knowledge bases.

Neil McIntyre [8] have done a summary of the methods used now for computer-generated stories. Numerous systems have been created to far, with varying approaches and different data sources utilized. They pay particular attention to systems that use narrative grammars, story schemata, autonomous agents, problem-solving skills, and common sense to construct tales. The current state of work in natural language generation (NLG) is then covered in the second half of the chapter. In particular, the approaches that produce trainable components—content selection, sentence planning, document planning, and surface realization—in or-

der to partially implement the NLG pipeline as outlined in Rater and Dale (2000). Finally, they make the case that a narrative generating system ought to ideally incorporate elements from each of these domains. Then Neil McIntyre introduces their story generation system, designed to create short stories targeted at young children. This system is made up of trainable parts and operates from beginning to end. They go into great depth about each of our components for surface realization, phrase preparation, and content selection. In order to identify the finest sentences and tales, they define a generate-and-rank approach and structure the narrative generating assignment as a search issue. Models for assessing tales are trained using shallow document characteristics. They present an interesting model that was trained using ratings of Aesop's stories that were obtained from human respondents. They also justify the use of an entity-based local coherence model (Barzilay and Lapata, 2008). Lastly, the generate-and-rank system was assessed by contrasting it with two straightforward baselines using a human assessment research. They present a graph formulation for encoding the action progression of an individual story protagonist across a collection of texts.

By combining the graphs of two protagonists, story plots are created from the regions in which they interact. They show that the plot graphs have an impact on the stories the system is capable of creating, specifically with respect to content selection. They specifically introduce and advocate for the use of a genetic algorithm to optimize stories produced from plots. Because they enable the algorithm to more imaginatively explore a greater area of the narrative search space, GAs are ideally suited for this role. They develop crossover (recombination) and mutation operators that are unique to our narrative production job and provide a range of potential fitness functions that take local coherence into account. Their human assessment research, they compare the GA-based system with the generate-and-rank method and two baselines before concluding. They then present an additional feature that uses a database of commonsense knowledge facts to build upon the stories that were created (Singh, 2002). These information about tale entities and

their motivations for actions are provided, together with an explanation of the actions' results. The investigation of the system's portability takes up the second part of the chapter. They investigate the system's capacity to produce stories from a fresh corpus that represents a different domain. They also examine the system's capacity to carry out a novel function, namely, the completion of incomplete or partial tales. Overall, Neil McIntyre have shown that their system provides an excellent platform for developing extensions to improve more in future nod now their accuracy is 57%. the quality of the generated stories and for use in new tasks and domains.

Matthew Paul Fay(2014)[9] have been attempting to advance computational narrative comprehension. This work's integration of imagination through tale creation is a key component. Researchers in the discipline have been interested in computational tale generating methods since the 1970s, when two of the first narrative generation systems were developed. Matthew Paul Fay(2009) talked about the development of creative tale writing methods across time. Above all, he pointed out some of the advantages and disadvantages of different strategies. He describes each system's methodology for computational creativity and give a synopsis of its main attributes. Character modeling plays a major role in his thesis, thus he also showing how it has been applied to the creation of stories in the past. By doing this, he provided the groundwork for how I will advance state-of-the-art computational narrative interpretation via character modeling and story production. Generally speaking, almost all story generating systems developed to date may be divided into three unofficial categories: world models, tale models, and author models (Bailey, 1999). The author models make an effort to mimic human writers' techniques. Rule grammars and other structural narrative representations are used by the story models. The world models try to establish a starting point and simulate ahead from there. The world modeling area would be where technique mostly fits in because of its emphasis on character modeling and simulation. The research gained overall accuracy of 62%.

## 2.4 Work In Mandarin

Henglin Huang and Chen Tang and Tyler Loakman and Frank Guerin and Chenghua Lin's [10] story generating problem is developed based on the Chinese story generation benchmark. The task is defined as follows The input is an outline X that has an unordered list of any number of Chinese phrases pertaining to characters and events. For the model to produce a cogent narrative Y is equal to $y_1, y_2, ..., y_n$, where $y_i$ is the story's i-th token (a Chinese character). They have used HanLP (He and Choi, 2021) to interpret Chinese tales' dependencies. Chinese dependency parsing uses the word segment, represented as $Seg = token_1, ..., token_m$, which comprises m tokens, as the fundamental unit, unlike English dependency parsing. As a result, the representation of a tale is $Y = Seg_1, ....$

To introduce target labels Ttarget into the original tales, they first identify the collection of dependencies $T = Segh, Dtag, Segt$ for each narrative. According to Henglin Huang and Chen Tang and Tyler Loakman and Frank Guerin and Chenghua Lin's these target labels are nsubj (representing subjects), root (typically representing verbs), dobj (representing direct objects), and pobj (representing indirect objects after prepositions). They provide a novel architecture for Chinese narrative production that consists of a neural conditional generator, a semantic denoising module, and a dependency tagging module. Their goal is to make Chinese generation better by better using the dependencies and semantics aspects. These traits greatly aid in the process of Chinese tale production, as seen by the performance increases observed in their tests and ablation studies. They gained over all accuracy of 67%. They examine several earlier studies on sentiment analysis using deep learning models, such as Word Representation, Sequence Models, and Convolutional Neural Networks. Text-to-text and data-to-text generation are two categories under NLG. Text-to-text generation is further subdivided into text correction, machine translation, summarization, and simplification, interpreting texts, formulating queries, etc. Lin, Jhe-Wei and Chang, Rong-Guey[11] used

statistical techniques to translate text in the realm of machine translation. They provided an approach for estimating these model parameters for a set of mutually translated sentence pairs, along with descriptions of five statistical models of the translation process. Although they only provided instances of translations between French and English, they thought the concept would also translate well across other language pairs. In 2003, Och al. employed statistical or heuristic models for common models to show and analyze several word alignment calculation techniques.

Lin, Jhe-Wei and Chang, Rong-Guey extracted and produced paraphrases using a multilingual parallel corpus. They demonstrated how to use a phrase in a different language as a pivot to identify paraphrases in a different language using phrase-based statistical machine translation alignment technology. They described how to modify it to take contextual information into account and created a paraphrase probability that enables the interpretations taken from a bilingual parallel corpus to be prioritized using translation probabilities. Regardless of the surrounding context, solitary sentences are typically used for abstract generation. In 2010, Clarke et al. presented a methodology for informative and cogent document compression. Their approach was developed using the framework of integer linear programming and was influenced by local coherence theory. According to the experimental findings, their model performed the best at the time. In 2010, Bartoli et al. released a study describing a technology designed to produce fictitious evaluations of scientific papers automatically. The tool's foundational feature is that it draws from a limited body of information. Naturally, another crucial area of NLG study is producing text from non-text data. The distinction between short-term and long-term memory in the human brain's memory mechanism—a variation of the artificial neural network, served as the basis for the Long-Short Term Memory Network (LSTM) model. It was employed to create a language model (LM). The popularity of deep learning has led to more modifications to LSTM-based deep learning models. To develop machine translation, a standard sequence-to-sequence

model (seq2seq model) was suggested. This paradigm is specifically designed to handle data with strings as both the input and the output. According to Lin, Jhe-Wei and Chang, Rong-Guey, the Seq2seq model makes use of two LSTM models: the first encodes the input sequence into a context vector, and the second decodes the context vector into an output sequence. A context vector is inserted between two LSTM models by this model. The input sequence's semantic meaning is represented by the context vector. The neural network can learn even in cases when there is a discrepancy in length between the model input and output sequences. GAN has been used in the field of NLP by several academics. Due to its incapacity to handle discrete data, the classic GAN architecture cannot be directly used, and related research has been at a standstill for the past three years. Lin, Jhe-Wei and Chang, Rong-Guey suggested SeqGAN as a solution to the issue of classic GANs' inability to handle discrete data by using Policy Gradient. Numerous academics have improved upon the SeqGAN design, and publications on the use of GAN in NLP have been published. However, no study on GAN has suggested using the abstract as an input and the article as the output model. He gained over all accuracy of 69%.

According tp Lin, Jhe-Wei and Tseng, Jo-Han and Chang, Rong-Guey[12] To handle data in sequence-to-sequence pair formats, the standard sequence-to-sequence (seq2seq) model is suggested. A seq2seq model encodes the input sequence to a context vector and decodes this vector for the output sequence using two deep Long Short Term Memory Networks (LSTM). Following the proposal of the seq2seq model, the attention mechanism is used. Compared to conventional seq2seq, seq2seq with attention is more efficient. Thus, a key element of the seq2seq paradigm is the attention mechanism. This approach may be used to generate poetry in Chinese. have been published, but no study on GAN that uses the article as the output model and the abstract as an input has been suggested. Models of deep neural networks have been applied extensively to abstractive text summarization. These days, convolutional sequence-to-sequence (ConvS2S) is a use-

ful method for summarizing abstractive material. Although convolutional neural networks perform exceptionally well in natural language processing (NLP) tasks, attentional processes remain an important component. On the other hand, text summarization produces the article's output sequence. The goal of neural story creation is to produce a narrative based on user-defined scenarios. Additionally, writers would employ as many phrases as they could to illustrate a point. As a result, gradient vanishing issues in RNN-based models will get worse, while memory leak issues may arise in CNN-based models. Consequently, gradient vanishing issues will worsen in RNN-based models, while memory leak issues may arise in CNN-based models. Another use for natural language processing (NLP) activities is question answering, where a computer can respond to queries in dialogues. Several models attempted to include it. The field of question answering has a wealth of resources. For instance, SQuAD's original edition included more than 100,000 questions, however its second version had questions that couldn't be answered in order to enhance the dataset's quality.

Regretfully, traditional Chinese cannot use SQuAD; it is just for English. When creating a tale, a series of runs of question-answering might be used to generate a lot of conversation. Image creation is successfully accomplished using generative Adversarial Nets (GAN) and conditional Generative Adversarial Nets (conditional GAN). GAN is the foundation of several generation models, including Cycle GAN. This method is also used in natural language processing (NLP) activities to address open-domain text generation issues that seek to simulate the sequential creation of discrete tokens. On the other hand, GANs emphasize an entirely automated approach that doesn't require any prior knowledge. A model to address neural narrative production was presented by Facebook AI Research, which is another significant use of ConvS2S. It developed a self-attention mechanism and a hierarchical model that represents the input sequence and interacts with proximate input items at lower layers and distant elements at higher layers. They gained overall accuracy of 61%.

## 2.5 Work in French

This paper study presents a complete approach that blends artificial intelligence (AI), sophisticated deep learning, and language norms to generate natural-sounding French sentences[13]. The foundation of this method is the Generative Pre-trained transformer (GPT) model, which represents a major breakthrough in creating text that resembles French written by humans. During the pre-training phase, the model is exposed to a wide range of French texts, including literary masterpieces and internet resources. The model has to understand the basics of the French language, including syntax, vocabulary, and stylistic nuances, in order to build a strong foundation for more complex language processing. After undergoing its first training, the model learns to anticipate the next word in a sentence by mastering context and semantics. This ability is essential for creating logical and contextually relevant sentences by following the natural flow of the French language. With datasets focused on certain themes or lexicons, the model can be further refined to align the sentences with specific needs. This flexibility makes the approach more applicable in a wider range of fields. A user-defined prompt is used to initiate sentence production, emphasizing the interactive aspect of the model.

By breaking up the prompt into smaller components or tokens, the model makes it easier to form sentences in an organized manner. Subsequently, it anticipates the subsequent tokens to construct a logical phrase, utilizing acquired grammatical and syntactic principles to guarantee compliance with French language conventions. Following the generation of the sentences, a post-processing stage could improve the sentences' tone, style, and fluency to make sure they are understandable to human readers. By receiving feedback from human assessors, the model's performance is continuously improved, increasing its ability to generate phrases that are genuinely human-like. This approach emphasizes how important AI and machine learning are to the advancement of natural language generation and processing. The ongoing development of these technologies makes it possible to produce text

that is more and more like writing by humans, not only in French but also in other languages. This technology has a wide range of possible uses, from improving digital communication to automating content creation. It also opens up new possibilities for inclusive and interesting language technologies.

## 2.6 Work In Arabic

Their paper Arabic Poems Generation primarily focuses on the creation of Arabic poetry[14]. One could think of creating stories in Arabic. First, in this case, the creation of poems begins with pre-processing the data to clean it up for the best possible model training. Here, they discarded any unnecessary punctuation, symbols, and letters that don't add anything to the sense of the Arabic text. Make that the models are trained on appropriately formatted, pertinent data after cleaning this dataset. They first method used in this case is the character-based LSTM model. It is implemented using the Keras toolkit and has a neural network structure. An embedding layer that quickly transforms characters into 256-dimensional vectors is the first step in this LSTM model. For this reason, every character in the dataset is mapped to a distinct integer index. And here, an LSTM layer with 1024 units comes after the embedding layer. An important factor in managing data sequences is Long Short Term Memory (LSTM), which is especially helpful for text. This layer aids the model in resolving issues like as the gradient problem and capturing long-term dependencies in the text. The last layer is thought of as a dense layer that foretells the results of the subsequent layer. The output layer's size matches the vocabulary size in addition to capturing each individual character. The 200-character input sequence is sent into the model, which uses the sequence to predict the following character. In this case, the prediction procedure is similar to a classification problem in which we see that every single character belongs to a class. Additionally, this research uses the category cross entropy loss function to optimize the model. The model compilation and effective training process are enabled by the selection of the Adam optimizer. Although the train-

ing process uses sequences with a predetermined length, this methodology accepts input strings of any length. Character by character, an example produced with this LSTM model shows that it can produce cohesive text. The Markov LSTM model, which combines the Markov model with LSTM technology, is the second model used here. Here, the Markov model uses the current word to predict the likelihood of the subsequent word.

The model of Arabic Poems Generation begins the poem with a word that is selected at random. then chooses the following word based on probability to maintain thematic and logical consistency. Words are selected to ensure a coherent flow of thoughts based on how likely they are to appear. Some words are not included in the forecasts in order to accommodate cultural sensitivities. An LSTM model is then given the verses produced by the Markov model. The four layers of this LSTM model are intended to improve and expand on the original verses. It makes use of the LSTM's long-term dependency memory, which is essential for preserving poetic form. The goal of the LSTM model is to forecast the subsequent verse's rhyme and meter. The model produces whole poems that follow Arabic rules by repeating this process. An innovative method for automatic poetry generation is presented by the combination of Markov models and LSTM. This process guarantees that the poetry produced are meaningful in addition to being good technically. Based on a dataset that has been specially pre-processed for this task, the models are trained and optimized. This methodology leverages the advantages of both statistical techniques and neural network models. Text creation at the character level is a strong suit for the Character-based LSTM model. On the other hand, the Markov-LSTM model guarantees structural coherence and theme relevance. Each model makes a distinct contribution to the production of Arabic praise poems that highlight the collaboration of several AI techniques. Poetic convention observance, coherence, and language quality are assessed in the produced poems. Especially when considered in the context of Arabic literature, this methodology constitutes a noteworthy advancement in the

field of computational creativity. They can automate the creation of Arabic poems with rich linguistic and cultural content by utilizing cutting-edge artificial intelligence technology. This creative method offers up new directions for investigating the relationship between technology and conventional literary forms, giving the current generation of Arabic poets a new outlook.

Another Arabic paper their research explores the potential of artificial intelligence to produce Arabic poetry through a combination of technological innovation and in-depth cultural knowledge[15]. To provide training data for AI models, the initial stage entails compiling a sizable dataset of Arabic poetry that contains both traditional and contemporary pieces. Particular machine learning methods that are well-known for managing problems related to natural language processing are selected. The data must be preprocessed, or cleaned up and arranged so that the models can interpret it, before training can start. In order to do this, the poetry must be divided into smaller sections, such as verses and lines, and then encoded in a way that makes sense for machine learning. Their tests use a range of machine learning models, from more sophisticated deep learning networks to more conventional approaches. These models are fed prepared data to train them, enabling them to pick up knowledge from the structures and patterns seen in Arabic poetry. Training is a difficult procedure that needs to be closely watched and adjusted to make sure the models are picking up the necessary skills. Enabling the models to produce fresh poetry that is close to the caliber and style of the training samples is the aim. Following training, the models' capacity to write poetry is evaluated by providing them with a starting point and allowing them to produce lines using the knowledge they have gained. Not just any poetry, but poetry that is authentic to Arabic literary traditions and feels right at home, is the aim. In order to achieve this, the models take into account the unique rhythm and structure of Arabic poetry. The produced poems are assessed by contrasting them with the norms of traditional Arabic poetry, taking into consideration elements such as emotional nuance and word choice. Poetry specialists in Arabic

evaluate the poems the models produce and offer commentary on the degree of resemblance between the computer-generated and human-written poetry. This feedback is essential for modifying the models so that they produce poetry more effectively. Additionally, software tools are employed to automatically assess the poetry according to a set of criteria, such as the poem's coherence, language use, and poetic device presence. The models are further adjusted in light of the evaluations and comments received. This testing, feedback-gathering, and adjustment procedure is performed multiple times. The last stage is to submit a selection of the greatest poetry the models wrote to the general public and experts alike, seeing how they are received. Many of them are startled that a computer can produce such poetry, but their reactions have been overwhelmingly positive. This research demonstrates the possible applications of artificial intelligence in poetry and other creative domains. It also invites debates about the definition of creativity and the viability of artificial intelligence. Overall, the concept encourages more research into artificial intelligence in the arts by fusing sophisticated machine learning techniques with a profound understanding for Arabic poetry.

## 2.7 Work In Spanish

The first paper effort, Natural Language Generation (NLG) is used to automatically create Spanish text, demonstrating a sophisticated marriage of computer techniques with language nuances[16]. Their work presents a complete system that combines extensive lexical databases with sophisticated syntactic analysis to produce a multi-layered approach to text production. This methodology relies heavily on the aLexiS lexicon, a vast lexicon of linguistic resources that have been meticulously developed to capture the nuances of the Spanish language. The lexicon adheres to strict standards and is based on foundational materials such as OSLIN-es and the Lexicon of Spanish Inflected Forms. Maintaining consistency throughout the dataset during this step is crucial as it facilitates the subsequent stages of verification and merging. Ensuring the authenticity and relevance of the

lexicon's content requires verification against the authoritative Diccionario de la Real Academia Española, which is a crucial stage in the process. The procedure contributes to the lexicon's continued high caliber, making it a solid basis for text production. By automatically selecting verb-preposition pairs from an analysis of Spanish literature, the lexicon is progressively enlarged, improving its depth and versatility in various NLG settings. In order to describe the intricate syntax of Spanish, define-clause grammar (DCG)—a method of syntactic structuring—is essential. In order to produce text that is both grammatically correct and syntactically coherent, this allows for the dynamic generation of sentences with verbs correctly conjugated to meet the surrounding syntactic context. Three stages comprise the autonomously operating system architecture for NLG: Text Planner, Sentence Planner, and Realizer. The Text Planner arranges input words into a logical structure at the beginning of the process, usually in the Subject-Verb-Object order. After that, the Sentence Planner polishes the structure by adding the required grammatical components. The last touches are made by the Realizer, who ensures that the sentence conforms to linguistic elements like gender, number, and person and modifies it for proper morphology. The system's flexibility in recognizing sentence structures and smoothly including subjects when necessary is one of its strong points. It can also adjust to the subtleties of Spanish grammar. The system effectively examines and chooses the most relevant syntactic structures using depth-first search techniques, guaranteeing accurate and fluid generated language. To sum up, this paper's methodology, which combines a thorough syntactic analysis, an organized generation process, and a rich lexicon, marks a substantial improvement in NLG for the Spanish language. This all-encompassing method offers a scalable and practical solution for automated Spanish text synthesis, which not only raises the caliber of the produced content but also makes a significant addition to the field of natural language generation.

Another research presents a methodology that combines linguistic resources and domain-specific ontologies to generate Spanish text from a chemical knowledge

base[17]. Using the METHONTOLOGY framework, the method entails arranging chemical data into a structured ontology that gives the system access to a clear set of chemical ideas and linkages. The text generation process is based on this organized knowledge base, which guarantees that the produced texts cover the entire chemical domain. In order to close the gap between domain-specific information and the language phrases required to produce Spanish writings, the Generalized Upper Model (GUM) is utilized. GUM plays a crucial part in the creation of semantically rich and grammatically sound Spanish phrases by classifying meanings and grammatical structures.

The ontology has shown adaptable and efficient in natural language generation projects across multiple languages, as seen by its successful adaptation from applications in English, German, and Italian. Specific linguistic traits of the Spanish language, like gender and number agreement, were carefully taken into account when adapting GUM for Spanish. In order to make sure that the produced texts correspond with the syntactic and semantic subtleties of Spanish, the adaptation procedure entailed examining the lexico-grammatical differences between Spanish and the languages that GUM had previously modeled. Additionally, the development environment KPML (Komet-Penman Multilingual) is used to create Spanish grammatical materials. KPML offers a framework for the creation and administration of linguistic resources, which facilitates the production of texts in different languages. The process of translating KPML to Spanish required a thorough analysis of Spanish grammar, with an emphasis on recognizing and applying the grammatical norms and restrictions that control the language. Choosing representative texts for analysis, comparing them to the English grammatical resources already in existence, and then modifying or developing new specifications for Spanish were all steps in the creation process of KPML. This made sure that the grammar resources for Spanish were complete and able to assist in producing documents that were both coherent and acceptable for their context. Their paper's methodology emphasizes how crucial it is to combine linguistic and domain ontologies for

efficient text production. The system produces comprehensible and educational writings in Spanish by utilizing the organized information offered by the chemicals ontology and merging it with the language skills of GUM and KPML. This shows the possibility for multilingual information sharing and reuse in addition to making chemical knowledge more accessible to a larger audience. Future additions of new languages or domains are made possible by the system's easy updates and expansions due to its modular architecture. One significant benefit is its versatility, which allows for the system's application not only in the chemical realm but also in other fields of study. In conclusion, this work offers a thorough approach for gene. Spanish texts are rated using a chemical ontology and the METHONTOLOGY framework, along with GUM and KPML. By guaranteeing that the produced texts are accurate in language and instructive, the method offers a useful resource for obtaining chemical knowledge in Spanish. This approach not only improves chemical information accessibility, but it also establishes a standard for future advancements in multilingual text generating systems.

## 2.8 Work In Hindi

In first their Hindi research paper, creating headlines for Hindi news items is a challenging task that combines sophisticated computer methods with knowledge gleaned from an extensive analysis of previous research[18]. The process starts with gathering a wide range of news stories from reputable Hindi news outlets, guaranteeing that a wide range of subjects and writing styles are covered. Thorough data cleaning is the first stage in preparing the content for future processing. This includes removing HTML tags, emoji, and superfluous spaces. Data standardization is accomplished through the use of tools like Genism and iNLTK, which lay the foundation for efficient model training. The models that are used for fine-tuning—someman/BART-hindi, facebook/mbart-large-50, indicBART, and mT5—are crucial to this process. These models are chosen based on their demonstrated skills to comprehend and produce language, and they are refined using datasets gathered

especially for the purpose of creating Hindi headlines. The models are adjusted through this procedure to better handle the subtleties of the Hindi language. The models learn how to create succinct, informative headlines that resemble those written by human writers by repeatedly training them on pairs of news items and their headlines. These refined models' efficacy is put to the test with a battery of measures, including ROUGE, BLEU, BERT scores, and semantic similarity scores, to make sure the models can accurately mimic the language and style of headlines created by humans. The algorithms are regularly improved to generate headlines that adhere to journalistic standards and successfully captivate readers through iterative fine-tuning and evaluation.

In the paper, their approach takes advantage of the capabilities of state-of-the-art NLP models and adapts them to the particular job of Hindi headline development. This not only promotes automated text summarization but also emphasizes the significance of varied datasets, the vital role of model selection, and the tactical fine-tuning of models to get better results. This is study methodology is scalable and replicable, with the goal of expanding its application to additional languages and NLP domains. By providing practical approaches to addressing the complexities of Indic languages, it will make a noteworthy impact on the field. A comprehensive approach to the problem is ensured by using a variety of models with different strengths, and a close attention to qualitative and quantitative evaluations of model performance provides profound insights into the efficacy of the process. This approach, which is transparent and repeatable, promotes additional NLP research and innovation. In the end, this research fosters tolerance and diversity in AI while also expanding the toolkit accessible for NLP applications in Indic languages. It successfully closes the gap between theoretical study and real-world journalistic application, demonstrating AI's revolutionary ability to meet the demands of Hindi-speaking consumers.

Creating Hindi writing that makes sense and fits the context well is a major chal-

lenge in this research on computational linguistics. Their paper offers a solution to this problem that combines novel machine learning techniques with language rules. It begins with a thorough analysis of Hindi's structure and meaning, drawing on a variety of sources to create a comprehensive language model. Getting a large collection of Hindi writing from all genres and locations is the first stage. This compilation is essential for upcoming research and model training in an effort to fully represent the vast diversity of Hindi. Once gathered, the data is cleaned to remove non-English characters, site codes, and emoticons to improve the text for analysis. Model training is more successful as a result of this cleaning, which provides a more accurate image of Hindi language patterns.

The text is then divided into smaller units, such as words and sentences. For a thorough analysis of the language's structure, this breakdown is essential. Each item gains an additional layer of analysis when part-of-speech tags are added, which enhances comprehension of Hindi grammar. N-gram models are useful for word prediction, which is necessary to create well-flowing, contextually appropriate writing. The significance of the text is deepened by word embedding, which aids in understanding the nuanced applications and meanings of words. Text creation relies heavily on neural networks such as RNNs, LSTMs, and GRUs because of their ability to handle sequences and produce coherent, flowing text. Attention mechanisms maintain consistency and relevance in the output by concentrating on the key components of the input. Transfer learning reduces the requirement for a large amount of new data and speeds up the model's ability to understand complicated patterns by utilizing knowledge from previously trained models. Ensuring the text is authentic and complies with language standards is a continuous process of improvement. It is important to get feedback from language specialists and Hindi speakers in order to modify the model and incorporate the cultural and stylistic details of Hindi[19].

## 2.9 Conclusion

By reading those papers for different languages we got step by step instructions to proceed with our research. Here we have an idea of how to collect data and how to store that data. They used the text dataset for their research where some have developed their own datasets for their research while others have used open source datasets. Then we learned how to process the data, learned about removing unnecessary words, symbols, tokenization, normalization and numbers from datasets using various data pre-processing codes. Here the dataset transformation has to be trained based on the model. After train data training, apply different models of machine learning and deep learning algorithm and do the generation process.

# Chapter 3

# Methodology

## 3.1 Introduction

Generation of story or text is generating coherent and meaningful text based on user prompt or the previous words. In our analysis of story generation of Bengali text, includes various algorithms that are generate stories and we select more appropriate model from this used models. Before training process of models, we done many pre-processing technique to clean our dataset and make our data organized like model that want but this processes take more time for not exists sufficient previous works in Bengali story generation. Some of techniques are irrelevant words and punctuation marks remove. This all of process in our works we have bagged this methodology part into 7 parts. Section 3.2 is our system architecture where it visualize our all processes briefly, and section 3.3 discuss details about our dataset and its features. In section 3.4 we discuss our Bengali wordcloud. Section 3.5 discuss preprocessing part of our dataset. In section 3.7 we discuss our applied algorithm and techniques then in the same Section Algorithm 1 we show our proposed algorithm. Finally, Section 3.8 add a conclusion to end up the chapter.

## 3.2 System Architecture

Figure 3.1 - Represent our proposed architecture that shows overall processes to get more expected results in our research.

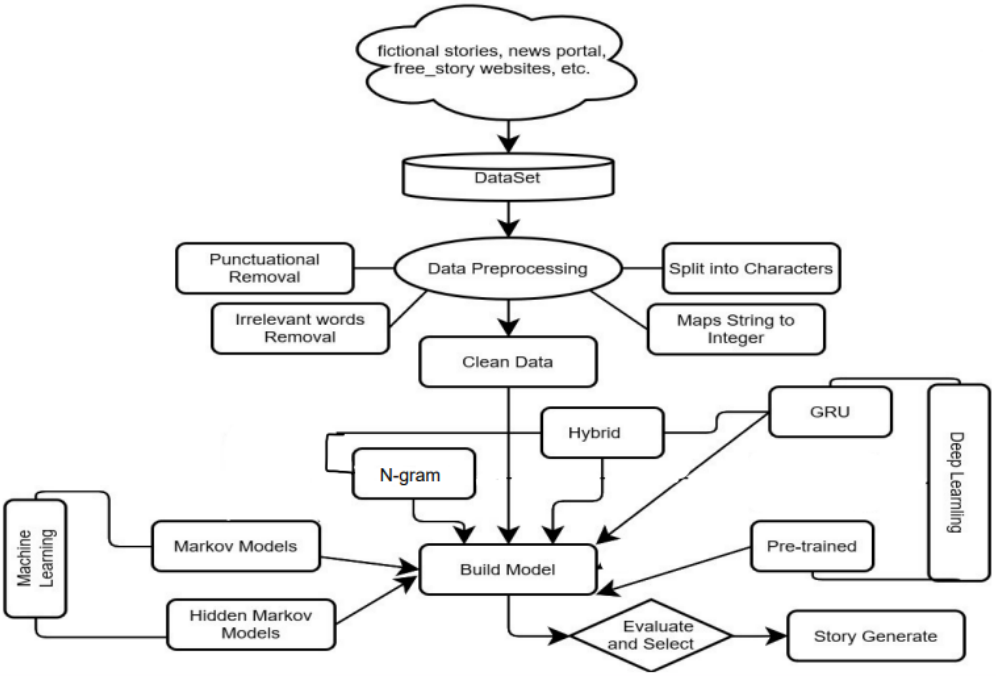


Figure 3.1: Our Proposed Architecture of Bangla Story Generation

## 3.3 Dataset

Bangla story sources and pre-built datasets are not available, and due to this, we take more time to collect more suitable stories. In traditional stories use classical Bangla language, idioms, and proverbs, but modern stories reflect the use of contemporary language, slang, colloquialisms, and borrowings from other languages. So here we cannot use the two jointly and have dealt only with modern stories. In this case, the number of stories is reduced to the maximum number of stories in Bengali. We collected a total of over 3,100 stories from various sources and built a dataset for our models. Our dataset has 4 columns Resource, Types, Name, and Story. Fig. 3.2 - Show our Dataset.

| | Resource | Types | Name | Story |
|---|---|---|---|---|
| 0 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | সিংহ ও ইঁদুর | এক সিংহ তার গুহায় গভীর ঘুমে আচ্ছন্ন। হঠাৎ একট... |
| 1 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | শিকারী ও ঘুঘু | নদীর তীরের একটি গাছের উঁচু ডালে একটি ঘুঘু বসে ... |
| 2 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | যোগ্য পাত্র নির্বাচন | সুলতান ইবরাহীম বৃদ্ধ হয়ে পড়েছেন। বয়সের ভারে ন... |
| 3 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | উচিত জবাব | একবার এক নাস্তিক এক দরবেশের কাছে এসে চারটি প্র... |
| 4 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | একজন পরোপকারী অফিস প্রধান | আব্দুল হালীমের স্ত্রী রাবেয়া একজন বিদুষী, পতিপ... |
| 5 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | গুপ্তধন | জনৈক ব্যক্তি স্ত্রী, তিন পুত্র ও এক কন্যা রেখে... |
| 6 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | কৃপণ ও নিঃস্ব | অনেক দিন আগের কথা। আরব দেশে ছিল এক কৃপণ ব্যক্ত... |
| 7 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | পাঠশালা | আল্লাহ পাক সময় সময় দুনিয়ার ধন-সম্পদে বিভোর মান... |
| 8 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | খাদীজার পর্দা | খাদীজা অন্যান্য দিনের মত আজও খুব ভোরেই ঘুম থেক... |
| 9 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | পরিণামদর্শী ক্রীতদাস | জগতে যারা নিজেদের নাম অমর করে রেখেছেন, তারা কর... |
| 10 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | মানুষকে সন্তুষ্ট করার পরিণতি | এক ব্যক্তি তার ঘোড়ায় চড়ে সফরে বের হয়েছে। সাথে ... |
| 11 | প্লে ষ্টোর অ্যাপ | শিক্ষণীয় | সাড়ে তিন হাত মাটি | \r\nসুদীর্ঘ পথেরও শেষ আছে, আছে এই মোহনীয় বসুন... |

Figure 3.2: Our Collected Dataset on Bangla Stories

We have collected these stories from various sources where the maximum stories are taken from online story websites and social media like Facebook are the one sources that allow us to collect stories. Figure 3.3 - shows a Pie chart of different sources of collected stories of our datasets.

The sources from which we collected stories include online story websites like samayupdates, banglaparenting, ghumparanirgolpo, parobashiblog, etc. which present maximum story data. Other sources like social media, personal blogs, online news portals, and Play Store Apps are remarkable sources of our dataset stories.

**Online story websites:** Numerous stories created by authors with varying experiences, backgrounds, and cultures may be found on online story websites. This variety may add value to your dataset and offer a wide range of viewpoints, which is beneficial for a variety of study subjects. Stories in a variety of genres and

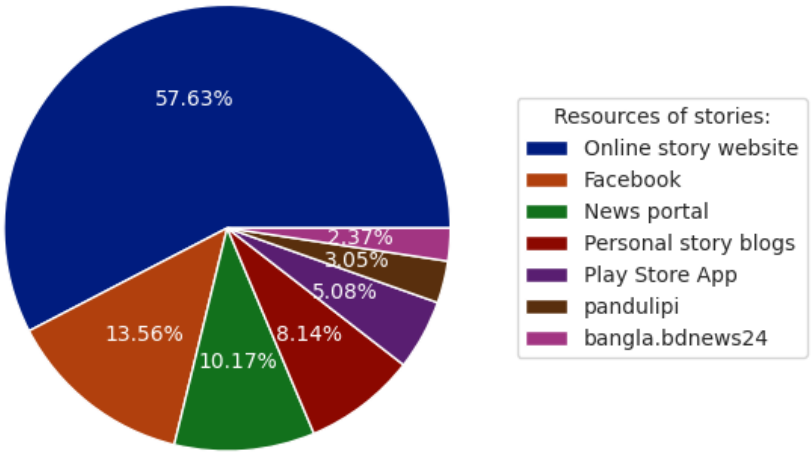


Figure 3.3: Pie chart of different sources of collected stories

types, such as fiction, non-fiction, fantasy, science fiction, romance, and more, are typically included on these platforms. Here anyone may record a wide range of language patterns, topics, and storytelling strategies because to this diversity. Online platforms provide narratives written in a conversational, organic style that mirrors individuals' creative expression. Understanding natural language use, including idioms, linguistic patterns, and colloquialisms, may be gained from analyzing this data. Narratives frequently mirror prevailing cultural ideals, attitudes, and concerns. You may gather narratives rooted in actual situations by compiling stories from internet platforms. This will increase the relevance of your study to current issues and conversations. When discussing delicate subjects, it might be morally appropriate to use fictitious narratives. Compared to tales based on actual events, these stories could represent less of a danger to privacy and secrecy because they are meant to be entertaining. Websites with online stories offer a handy and easily available source of information. It's simple to get a lot of text data without requiring a lot of resources or laborious data collecting procedures. Gathering narratives from digital platforms is frequently more economical than employing techniques like conducting interviews or surveys. It makes it unnecessary to find

volunteers and makes it possible to extract huge datasets quickly. Numerous study directions, including sentiment analysis, narrative structure analysis, genre categorization, and more, may be explored through story analysis. The abundance of data enables a broad range of analytical approaches and procedures.

**Facebook:** Facebook started its journey in Bangladesh in 2009. According to the information of Prothom Alo, there are currently about 47.2 million active Facebook accounts from Bangladesh. Which is about 28 percent of the total population of our country. From it is said that Facebook is very popular in the context of Bangladesh. There is a sizable and varied user base on Facebook. Gathering narratives from this site can offer a variety of viewpoints, experiences, and subjects, enhancing the richness of your collection. Facebook stories frequently depict real-world encounters, viewpoints, and exchanges. This can improve dataset's relevance and authenticity, making it more realistic and representative of daily life. Since Facebook is a social media site, tales shared there frequently deal with interpersonal relationships, social interactions, and community dynamics. Understanding how people express themselves and communicate in social situations might benefit from this. With Facebook's timeline function, you may chronologically analyze tales by gathering them over time. This can be helpful for monitoring patterns, linguistic alterations, or changes in the subjects that individuals are talking about. Because Facebook stories are primarily user-generated, it offers a chance to record natural language variations and expressions. This can be useful for training models that produce language that is both contextually relevant and human-like. Personal interests and experiences are frequently reflected in Facebook tales. Understanding customization in storytelling and adjusting produced stories to suit individual tastes may both benefit from this. Using Facebook content that has been posted openly can help allay ethical worries about privacy. Compared to using private or sensitive information, there is less chance of violating an individual's right to privacy because the material is already public.

**Personal blogs:** Personal blogs frequently have intricate and complex tales that provide a profound window into the experiences, viewpoints, and feelings of the individual. This depth might be useful for qualitative research that seeks to comprehend personal narratives in great detail. On their personal blogs, bloggers discuss a wide range of subjects, from wellness and health to travel and leisure. Because of this diversity, researchers are able to investigate a wide range of themes and topics in the context of individual narratives, which adds to the size of the dataset. So this is one of helpful source for our work to collect data.

**Online News Portal:** It is the online version of various newspapers where people can read or watch the news in news media such as newspapers or television channels. The online news portal is the online version of those mass media, where people get the benefit of reading the news as well as publishing their stories via creating their account on an online news portal site. These portals are newspapers, many media outlets publish their news only online. Nowadays, people have reduced the direct purchase or reading of newspapers. Because they offer current information, online newspapers are helpful to researchers who are looking to examine topics, trends, or events that are happening right now. Newspapers cover a wide range of subjects, including sports, entertainment, politics, and the economy. Because of this diversity, researchers can examine a broad range of topics in a single dataset. News organizations frequently uphold journalistic standards and have a solid reputation. Adding narratives from reliable sources can improve the data's dependability. Newspapers may offer several viewpoints on a certain topic like stories, enabling scholars to examine and contrast the ways in which various media portray and interpret the same events. Newspapers can serve stories as a significant resource for scholars examining the link between social ideas and media depiction, as they often reflect popular opinion. Most of the time they collect their required news from online news portals or visit the websites of news portals to read various news to save time. On the other hand, people take television as a

part of their entertainment, but due to lack of time, many people do not waste a long time in front of the television and only read some news from the websites of television channels. The most interesting thing is that many newspapers give the facility to write stories from users and it also helps us to collect stories from this online system of news portal.

**Play Store Apps:** Users may produce and share content, such as stories, reviews, comments, and feedback, using a variety of Play Store applications. By examining this user-generated material, analysts can learn more about the preferences, viewpoints, and experiences of the app's users. Real-world user experiences with the app are frequently reflected in Play Store stories and reviews. This data might be useful to understand how users interact with the app, what features they enjoy and don't like, and any problems they run into. Some apps combine narratives with multimedia components including pictures, videos, and sounds. If you are conducting research that requires you to analyze both textual and multimedia data, gathering such information may prove advantageous. User opinions and emotions are frequently included in reviews and stories on the Play Store. Stories may be analyzed over time to learn more about how the app changes, how upgrades affect it, and how users respond to those changes. This long-term viewpoint can aid in improving comprehension of the app's lifespan. Stories that have been gathered may be used to evaluate and contrast other apps in a certain category. This method can highlight opportunities for improvement as well as patterns and trends in a variety of applications.

Now, in the figure 3.4 shows the overall length idea of the collected dataset where the maximum story taken in length(count by words present in story) more than 7400 where the minimum length less than 130 words. From the generated bar plot of length, the average length is approximate 960 words and all are cleaned text.

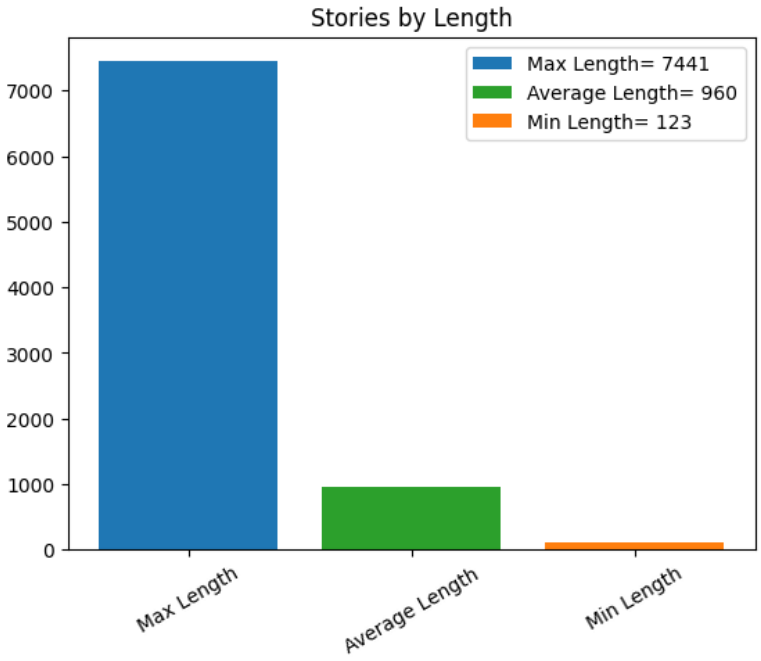


Figure 3.4: Distribution of story lengths for highest, average, and lowest for our collected dataset

## 3.4 WordCloud

WordCloud is a dynamic visual representation tool for text processing and is used here for Bengali text using different colors to represent the most frequent Bengali words. The process is also known alternatively as tag cloud or text cloud. In this process we used story names and stories for most frequent stories and word shows. Here, the size of Bengali words in our dataset with high frequency in the generated image is large, while the size of low-frequency words is relatively small compared to other words.

The wordClouds generated from story names and stories are illustrated in Figure 3.5 and Figure 3.6 accordingly. In this process we remove some stop words like "এবং", "সে", "গল্প", "এই", "ও", "সে", "তার", "তাই", "তাই", "তার", "তুমি", "আমার", "গল্প" , "না" , and etc. for shown wordClouds get more organize. Here we use regex for remove the punctuation and other symbols or text excluding Bengali text.


Figure 3.5: WordCloud for name or type of stories of our dataset


Figure 3.6: WordCloud for stories all frequent words

## 3.5 Pre-Processing

Data processing is very crucial before train the model with collected processed dataset. If the datasets are not cleaned the the output gets low accuracy and the training period will take lot of time. The total dataset data shows in figure 3.7. In the generation technique, it the dataset is not cleaned the generation text will

be unrecognized and irrelevant. Here are used pre-processing techniques of our model:

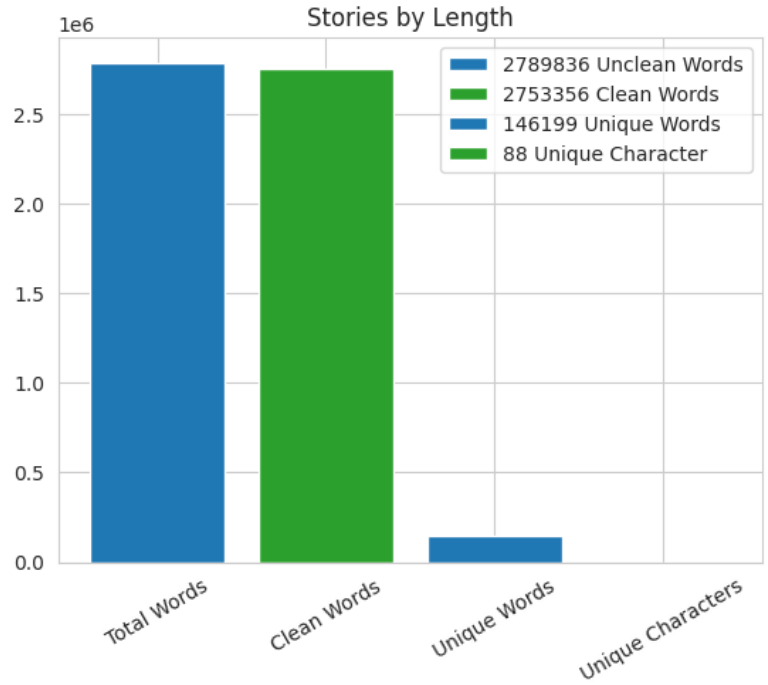


Figure 3.7: Bar plot of comparison of both unprocessed and clean data for both individual words and characters

- **Punctuation removal:** To clean the dataset, we firstly remove the extra punctuation marks present in our dataset. eg: (., „ :, ;, ”, ’, ?,<, >, -, &, @, #, !, *, ,  , $, %, ... etc). show in figure 3.8
- **Irrelevant words Clean:** In our dataset, there are many irrelevant words that have no meaning and contain words from different languages. English and Hindi were irrelevant characters for our dataset and first we select irrelevant characters and remove them from our data.
- **Tokenization:** To get good results the dataset needs to be tokenized which is a process that removes all types of spaces, newlines, punctuation marks and extra spaces from the dataset. For this, we used a manual code for tokenized dataset without using any library.

- **Drop Numbers:** This excludes all types of numbers from our dataset and it done manually. such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, "০", "১", "২", "৩", "৪", "৫", "৬", "৭", "৮", "৯" etc.

- **Encode words:** Encoding is very important in handling text data and communication between different systems. Here we split Unicode strings into substrings of individual characters and this is also useful for tokenizing text. Next we specify the encoding of the input string and set it to **UTF-8**, indicating that the input string is encoded using UTF-8.

- **Maps ids and words:** Mapping IDs to words (and vice versa) is a common task in natural language processing (NLP) and is particularly important for textual data for efficient representation as well as memory efficiency. In our system this is done using a vocabulary of text to get the unique ID.

| | name | story | story_ii | name_ii |
|---|---|---|---|---|
| 0 | সাত বোন ছোট পরী | চারিদিকে পাহাড় ঘেরা উপত্যকার মাঝে ছিল অপূর্ব স... | চারিদিকে পাহাড় ঘেরা উপত্যকার মাঝে ছিল অপূর্ব স... | সাত বোন ছোট পরী |
| 1 | সিনডারেলা | ছোটবেলায় কে না পড়েছি সিনডারেলার গল্প। যদিও বিভ... | ছোটবেলায় কে না পড়েছি সিনডারেলার গল্প। যদিও বিভ... | সিনডারেলা |
| 2 | প্রজাপতি রাজকুমারী | রাজকুমারীর মনে পড়ে গেল তার প্রতিজ্ঞার কথা। সে ... | রাজকুমারীর মনে পড়ে গেল তার প্রতিজ্ঞার কথা সে ব... | প্রজাপতি রাজকুমারী |
| 3 | \nহিমালয়ের ছোট্ট গ্রাম\n | সে অনেক কাল আগের কথা ,দূর হিমালয়ের কোলে ছিল এক... | সে অনেক কাল আগের কথা দূর হিমালয়ের কোলে ছিল এক ... | হিমালয়ের ছোট্ট গ্রাম |
| 4 | আলিফ লায়লার রাজ্য | ছোট্ট একটা রাজ্য। আরঙ্গান। তিন দিন ধরে একটানা ... | ছোট্ট একটা রাজ্য। আরঙ্গান। তিন দিন ধরে একটানা ... | আলিফ লায়লার রাজ্য |
| ... | ... | ... | ... | ... |
| 3069 | পর্ব-১৮: রাজকার্য ভুলে রাজা ধর্মের বাণী প্রচার... | মনে রাখতে হবে যে, রাজবাড়ির "সম্মার্জন কর্তা" হ... | মনে রাখতে হবে যে রাজবাড়ির সম্মার্জন কর্তা হলেও... | পর্ব রাজকার্য ভুলে রাজা ধর্মের বাণী প্রচারে ব্... |
| 3070 | পর্ব-১৯: উন্নতি করতে গেলে অপেক্ষায় বসে থাকলে চ... | আমাদের বিচিত্র এই জগৎ-সংসারে সকল মানুষই নিজেদে... | আমাদের বিচিত্র এই জগৎসংসারে সকল মানুষই নিজেদের... | পর্ব উন্নতি করতে গেলে অপেক্ষায় বসে থাকলে চলবে ... |
| 3071 | পর্ব-২০: দু'জন সন্ন্যাসী এক জায়গায় হলেই সাধন-ভ... | অতিভক্তি যে চোরের লক্ষণ সে কথাটা শিশুকাল থেকেই... | অতিভক্তি যে চোরের লক্ষণ সে কথাটা শিশুকাল থেকেই... | পর্ব দুজন সন্ন্যাসী এক জায়গায় হলেই সাধনভজন ভুল... |
| 3072 | পর্ব-২১: এ-কালের মতো সেই পুরোনো আমল থেকেই রাজন... | তাঁতির সেই বউটি ছিল একজন "পুংশ্চলি"। দেবদত্ত ব... | তাঁতির সেই বউটি ছিল একজন পুংশ্চলি। দেবদত্ত বলে... | পর্ব একালের মতো সেই পুরোনো আমল থেকেই রাজনৈতিক ... |
| 3073 | পর্ব-২২: কামুক পুরুষের সঙ্গ ছাড়া একজন নারীও এক... | অন্ধকাল গলিপথ দিয়ে নিঃশব্দে তাঁতি-বউ চলে গেল দ... | অন্ধকাল গলিপথ দিয়ে নিঃশব্দে তাঁতিবউ চলে গেল দে... | পর্ব কামুক পুরুষের সঙ্গ ছাড়া একজন নারীও একা কো... |

Figure 3.8: Cleaned data after removing all punctuation marks and other symbols or text

After processed all data the number of cleaned unique character is 64 and they are ' ', '।', 'ঁ', 'ং', 'ঃ', 'অ', 'আ', 'ই', 'ঈ', 'উ', 'ঊ', 'ঋ', 'এ', 'ঐ', 'ও', 'ঔ', 'ক', 'খ', 'গ', 'ঘ', 'ঙ', 'চ', 'ছ', 'জ', 'ঝ', 'ঞ', 'ট', 'ঠ', 'ড', 'ঢ', 'ণ', 'ত', 'থ', 'দ', 'ধ', 'ন', 'প', 'ফ', 'ব', 'ভ', 'ম', 'য', 'র', 'ল', 'শ', 'ষ', 'স', 'হ', '়', 'া', 'ি', 'ী', 'ু', 'ূ', 'ৃ', 'ে', 'ৈ', 'ো', 'ৌ', '্', 'ৎ', 'ড়', 'ঢ়', 'য়'.

## 3.6 Approach

We used a variety of machine learning and deep learning algorithms to analyze advanced generation methods. Models used are: Ngrams, HMMs, Markov models, RNNs and pre-trained models and algorithms.

### 3.6.1 N-gram

N-gram models are a fundamental concept in which it is a type of probabilistic language model used in natural language processing (NLP). It is a sequence of n contiguous words extracted from a text. Here we used bi-gram to generate sentences.

Calculation Formula:

Bigram Probability: $P(w_n|w_{n-1}) = C(w_{n-1}, w_n)/C(w_{n-1})$

where $C(w_{n-1}, w_n)$ is the count of the bigram (the two words occurring in sequence) and $C(w_{n-1})$ is the count of the preceding word in the corpus.

### 3.6.2 Markov model

When modeling randomly changing systems, a Markov model is a stochastic model in which it is assumed that future states depend only on the current state and not on the preceding sequence of events. The Markov property is the name given to this attribute. Markov models are used to model temporal sequences of events or states in a variety of domains, such as finance, economics, statistics, and physics.

### 3.6.3 Hidden Markov model(HMMs)

An assumption-based statistical model diagram in figure 3.9 in which the system under study is a Markov process with unobserved (hidden) states. In temporal pattern recognition applications like speech, handwriting, gesture identification, part-of-speech tagging, musical score following, and bioinformatics, HMM is extensively utilized.

Calculation Formula:

Transition Probability: $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$,

where qt is the state at time t, and $S_i, S_j$ are specific states. This calculates the probability of transitioning from state i to state j.

Emission Probability: $b_j(k) = P(v_k \text{ at } t | q_t = S_j)$,

where $v_k$ is the observed value at time t given the state $S_j$. This calculates the probability of observing $v_k$ given the current state is $S_j$.

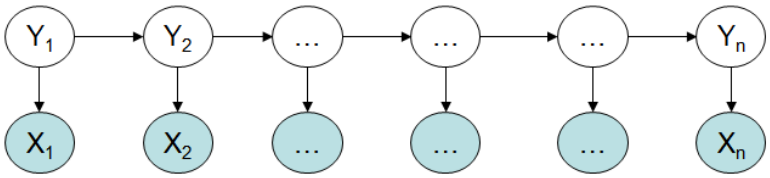


Figure 3.9: The diagram of Hidden Markov Model

### 3.6.4 Gated Recurrent Unit (GRU)

In recurrent neural networks, GRU(shown in figure 3.10) is a gating mechanism that was developed to address the vanishing gradient issue with conventional RNNs. With fewer parameters, it is comparable to an LSTM (long short-term memory) with forget gates.

Calculation Formula:

Update Gate: $z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$

Reset Gate : $r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$

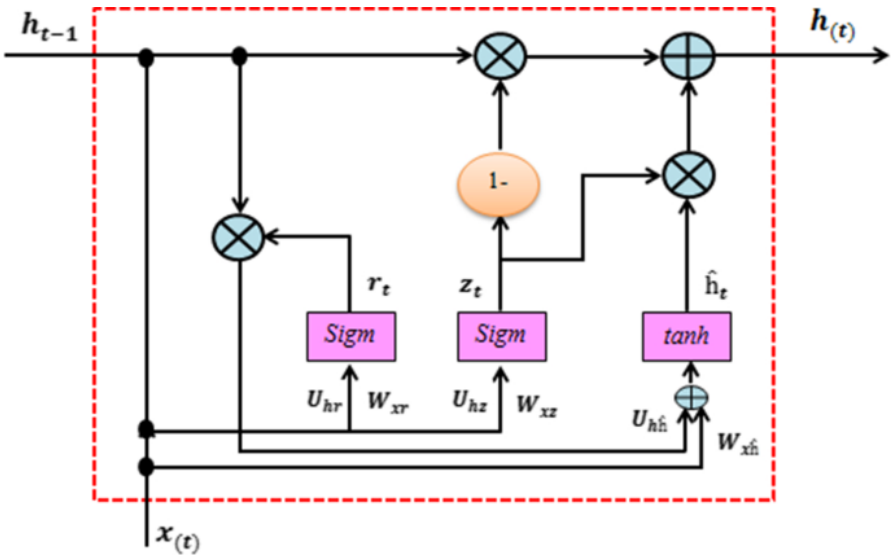


Figure 3.10: Diagram of the gated recurrent unit RNN

Candidate Activation: $\tilde{h}_t = \tanh(Wx_t + U(r_t h_{t-1}) + b_h)$

Final Activation: $h_t = (1 - z_t)h_{t-1} + z_t(h)_t$

### 3.6.5 Pre-trained model

Like RNNs and GRUs, the Transformer model where we can use pre-trained models is made to handle sequential data; however, it does not require sequential operations, enabling far higher parallelization. To depict global interdependence between input and output, it completely depends on self-attention processes. In our works, our main focus is not to use pre-trained models to generate a text while we only focus on story generation and one concern is language complexity. But here we try to generate text using pre-trained model on our dataset but it is still not compatible with Bengali story. There may not be that many pre-trained models available for Bengali as compared to English but bangla-gpt is one of the famous pre-trained models for Bengali text. There are not many pre-trained models available for Bengali as compared to English but Bangla-GPT is one of the known pre-trained models for Bengali text.

## 3.7 Our Proposed System

In our proposed algorithm, the stories will generate with best suitable combination of words where it compare both predicted words of GRU and N-gram algorithm and choice best option. In the given algorithm 1, shows the total procedures of input to output story generation and here used the transactional probabilities of words used for compare most suitable words combination.

**Algorithm 1:** Our Proposed Hybrid Algorithm

**Input:** User Inputs as a story prompt or keywords

**Output:** Generated story based on input

**Data:** $input\_Data = input()$

```
1  begin
2      current_word ⟵ input_process(input_Data);
3      Result ⟵ "";
4      sentences ⟵ 0;
5      while sentences < sentence_limit do
6          nextWord_1, state ⟵ GRUmodel.generate(current_word, state);
7          nextWord_2 ⟵ Ngram.generate(current_word);
8          if Prob(nextWord_1|current_word) >
             Prob(nextWord_2|current_word) then
9              current_word ⟵ nextWord_1;
10         end
11         else
12             current_word ⟵ nextWord_2;
13         end
14         Result.append(current_word);
15         sentences ⟵ count_sentence(Result);
16     end
17 end
```

## 3.8 Conclusion

The new approach is very important for our research to analyze the strategy to see if it is better than other generated texts in the model. Here pre-processing of data is crucial for efficient presentation as well as memory efficiency of the system. In the model training phase, we train all the models with cleaned data whereas the GRU model takes more time to complete all the epochs, the training of the pre-trained model also takes time. As a result, we allocated additional time for each training process to upgrade the dataset with new data or stories.

# Chapter 4

# Performance Evaluation

## 4.1 Introduction

Evaluating text or story generation models is a complex task with no best method and the evaluation method depends on the specific task and purpose of the model. Some of the approaches to evaluating story generation are human evaluation, automated metrics (BLEU, ROUGE, METEOR), consideration, and emergent methods[20]. In our research, we select human scores on the generated stories to get an estimated evaluation score. There are three main categories for evaluating our results of model and the terms are sentence structure or grammar check, story consistency as well as context or prompt or story theme.

### 4.1.1 Sentence Structure

Bengali language follows a **subject-object-verb (SOV)** sequence and constitutes a typical example of a Bengali language. So here we check the Bengali sentence structure and see if it follows the Bengali grammar which involves placing first the subject, then the object and finally the verb. Here is a breakdown of common sentence structures in Bengali:

- **Subject (S):** The entity performing the action or being described where it can be a noun, pronoun or noun phrase.

- **Object (O):** The entity that takes the action of the verb and is like the subject, the object can be a noun, pronoun or noun phrase.
- **Verb (V):** Indicates the action performed by the object and changes their form based on tense, aspect, mood and subject person and number in Bengali.

$$Sentence\_Structure\_Score = \frac{\sum_{i=1}^{n} Sentence\_Structure\_Score_i}{n}$$

where, $i$ indicate individual person and $n$ is total number of persons.

### 4.1.2 Consistency

It is important to convey information effectively, build credibility and improve the overall quality of written communication for all texts produced. In this process, ensuring consistency in spelling, grammar, punctuation, terminology, style and format allows writers to produce Bengali stories that are clear and coherent.

$$Consistency\_Score = \frac{\sum_{i=1}^{n} Consistency\_Score_i}{n}$$

### 4.1.3 Context

Considering the context of the Bengali text, it is essential for proper comprehension and interpretation, as it provides valuable background information and helps in understanding the purpose of the text as well as the story.

$$Context\_Score = \frac{\sum_{i=1}^{n} Context\_Score_i}{n}$$

The overall score calculated by,

$$Score = \frac{Sentence\_Structure\_Score + Consistency\_Score + Context\_Score}{3}$$

In all this process we collected 20 people's comments on the stories we created to evaluate the proposed algorithm.

| Output Results | | | | | |
|---|---|---|---|---|---|
| Statistic Measure | N-gram | Markov Model | HMM | GRU | bridNG |
| Sentence structure | 0.533 | 0.73 | 0.616 | 0.637 | 0.668 |
| Consistency | 0.34 | 0.517 | 0.15 | 0.32 | 0.423 |
| Context | 0.23 | 0.31 | 0.45 | 0.33 | 0.485 |
| Overall | 0.381 | 0.453 | 0.4053 | 0.429 | 0.5223 |

Table 4.1: Human and model Evaluated Generated Results of N-gram, Markov Model, HMM, GRU and bridNG on Dataset

## 4.2 Result Analysis

The proposed bridNG system gives better scores than machine learning and deep learning algorithms because it selects the best option from GRU and N-gram models. In the GRU model training phase it gives us some loss scores and in Figure 4.1 shows plot of this based on epochs vs losses. From the figure, it shows the loss rate in each epoch where the loss decreases as the number of epochs increases. In bridNG the stories model created, where sentence structure scores higher than other evaluative terms. Compared to other discrete models, it gives better generated scores.

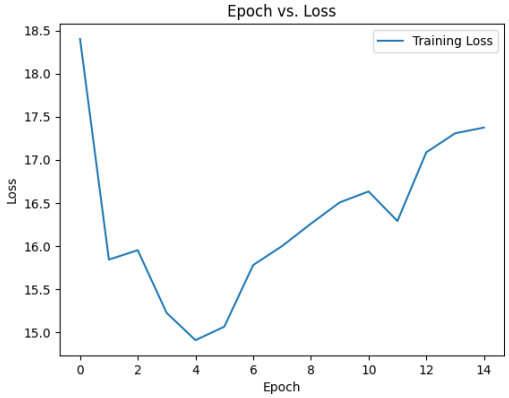


Figure 4.1: Epochs vs Losses plot of GRU model over our dataset

| **Paper Title** | **Approach** | **Data Set** | **Highest Accuracy** |
|---|---|---|---|
| **Our Paper:** Hybrid Approach to Automated Contextual Story Generation in Bangla | Hybrid Approach(ML & DL),GRU, Transformer, HMM, N-gram | We create our own data set which is more than 3100 from various online sources and majority data are collected from online stories website | The overall score of our proposed system is 52.23% |
| **Paper 1:** Towards Controllable Story Generation [21] | LSTM generation model, AMT,ROC-stories corpora | Gathereed 3980 annotated stories with the turker-researcher agreement | Estimated Accuracy of the model is 69%. |
| **Paper 2:** Strategies for Structuring Story Generation [22] | Semantic Role Labeling (SRL),seq2seq model, NER Entity Anonymization, byte-pair encoding(BPE) | 300k story premises paired with long stories and they focus on the prompt to story generation aspect of this task 1000 words and fixing the vocabulary size to 19,025 for prompts & 104,960 for stories. | SUMMARY 52%, KEYWORDS 43%, COMPRESSION 47%, SRL ONLY 54%, SRL+NER 57% SRL+COREF 60% |
| **Paper 3:** Calliope: Automatic Visual Data Story Generation from a Spreadsheet [23] | Monte Carlo tree search (MCTS),RNn | A total of 230 high-quality videos were selected and manually segmented into 4186 story pieces. | Overall accuracy 67% |

Table 4.2: The Comparison of text generation with other papers

## 4.3 Discussion

Generating stories for Bengali texts is a complex process where it is difficult to get the maximum evaluation score. The **bridNG** algorithm gives the highest score over all machine learning and RNN models (GRU) although the score is not very effective in generating relevant stories. Here, the system needs to be developed or trained with more than 100 thousand stories to get the best output. Less than 3500 Bengali stories are not enough to train the models to get the expected results or stories. Markov models and HMMs also produce stories where the Markov model can produce no more than 2 sentences on average but other models can produce the expected sentences. The proposed model uses probabilities to select the most frequent words that constitute the generated words of the stories where the GRU is the most costly to take up a lot of time in the training condition. Thus, the proposed method is more effective than other traditional models in our dataset.

# Chapter 5

# Conclusion

For Bengali narrative production, using both RNN (Recurrent Neural Network) and N-gram models offers a versatile method that blends the advantages of cutting-edge deep learning with conventional language modeling approaches. RNN provide a sophisticated way to produce a variety of imaginative and varied tales because of their capacity to maintain language subtleties and record contextual dependencies. The enhanced quality of created stories is a result of their capacity to adjust to the distinctive features of the Bengali language, including script and cultural nuances. However, N-gram models offer a useful starting point because of their simplicity and computational efficiency. Combining these models guarantees a thorough comprehension of long-range connections in sequential data and facilitates an adaptive and flexible narrative experience for the user. In the end, the combination of RNN and N-gram models promises to produce captivating, authentic, and culturally rich Bengali stories that will satisfy a variety of tastes and guarantee a dynamic story creation process. When endowed with long short-term memory (LSTM) cells, RNN are particularly effective in capturing contextual relationships in Bengali language input. This makes it possible to create stories that more effectively capture the subtleties and complexity of the language.

RNN are very good at comprehending and producing data sequences. Because it allows the model to keep a consistent tale by taking into account the context of the previous lines, this is advantageous for the creation of stories. Using a variety

of datasets for training, RNN can generate tales with a wide range of topics and genres. This adaptability is helpful when producing material for various audiences and tastes. When generating stories, RNN may be creative in bringing in fresh and original components. This may lead to stories that are distinctive and captivating, drawing readers in. Bengali is a unique language with its own phonetics, alphabet, and subtle cultural differences. These elements may be handled by RNN with customization, guaranteeing that the stories produced follow the language and cultural conventions of Bengali storytelling. Long-range dependencies in sequential data are intended to be captured by RNN. This aids in preserving coherence and consistency in the story even when events or details are spaced out by a large number of words in the context of storytelling.N-gram models can be used as a starting point for narrative generation even if they are not as advanced as RNN. They can give you a rapid start when creating simple stories and use less processing power.

A variety of language models may be created by combining RNN with N-gram models. Depending on the desired outcome, this diversity may be helpful in meeting various needs and preferences. The tales produced by these models are guaranteed to reflect the linguistic and cultural diversity of the Bengali language because they were trained on a sizable and varied corpus of Bengali language text. RNN enable a customized narrative experience by being adjusted in response to user input and feedback. User happiness and engagement are increased by this flexibility. It's important to note that model performance may vary based on factors such as data quality, model architecture, and training methodology. Experimentation and fine-tuning are crucial for achieving optimal results in generating high-quality Bengali stories. For our study on Bengali narrative creation, we have painstakingly assembled a large data set with more than 3000 stories. Both the RNN (Recurrent Neural Network) and N-gram models are trained and evaluated using this large data set as a basis. The data set captures the linguistic subtleties and cultural complexities of the Bengali language and is diverse, encompassing

a number of genres and subjects. By utilizing RNN capability, we want to capture the contextual dependencies present in the stories and guarantee a cogent and contextually appropriate tale generating process. In addition, the incorporation of N-gram models offers a comparative baseline, enabling us to investigate the trade-off between computational effectiveness and story complexity. Our research aims to promote Bengali narrative production with this extensive data set by providing insights into the interaction between contemporary deep learning techniques and conventional language modeling for a complex and culturally relevant storytelling experience.

Our research delves deeply into the relatively new and intriguing task of autonomously producing Bangla stories with computers. Bangla is a lovely language, but because of its intricate laws and patterns, it is difficult for a computer to comprehend it well enough to compose stories. Our aim was to use a clever combination of technologies to make this happen and contribute new stories to the literature of Bangla. Initially, we collected more than three thousand stories from a variety of sources, including websites, applications, and Facebook. This sizable collection was crucial since it provided a wealth of content for our computer model to study, encompassing a wide range of subjects and writing styles. This data needed to be cleaned up and organized before we could utilize it, which involved correcting mistakes and ensuring the language was in a format the computer could read. For this study, we primarily used a unique type of technology that blends GRUs with N-grams. GRUs facilitate the computer's retention and application of knowledge about the structure of ideas and narratives. It can comprehend and anticipate the next word based on previous words thanks to N-grams. They are a formidable team that can produce fresh and insightful Bangla narratives.

We wanted stories from the computer, but not just any stories, but good stories. As a result, we asked readers to evaluate the stories on the basis of their writing quality and coherence. This was an important milestone since it demonstrated how similar the stories produced by the computer were to those written by peo-

ple. It was evident from a comparison of our process with other approaches that ours produced more logical and entertaining stories. This achievement demonstrated that integrating GRUs and N-grams was a wise decision for addressing the particular difficulties posed by the Bangla language. To put it briefly, our study has expanded the realm of automated tale writing possibilities, particularly for Bangla. We've demonstrated that computers can produce fresh, captivating tales in this rich language when given the appropriate resources and techniques. This creates intriguing opportunities for the future, when we can witness even more inventive and sophisticated applications of technology in literature.

## 5.1 Limitation of the research

- The dataset we created has not more data on Bengali stories for unabailable data sources. But in this NLP and deep learning reasearch need to large amount of dataset.
- Our proposed bridNG algorithm not actually gives human like stories.
- Proposed approach is not more time effective for its more computational complexity for used of hybrid technique.
- Here used RNN model but in recent transformer based model are most famous.
- Our model not based on user input prompt where rule based algorithm is one of best for this.

## 5.2 Practical Implications

Generation of Bengali stories can have significant cultural, educational and entertainment impact where it will contribute to the preservation and promotion of Bengali culture and literature. The new stories generated will inspire our Bengali

traditional folktales, Puranas and literary works. The system helps maintain the richness of the Bengali storytelling tradition where it helps ensure our continuity for future generations. It serves as a valuable tool for language teaching and learning, especially for new language learners of Bengali.

Story Generation also helps in creating content for the Bengali entertainment industry as well as other audiences. Here it can create plotlines, characters and narratives for Bengali literature, films, television shows and digital media. It will be an interactive narrative experience in Bengali for a personalized storytelling platform. Users can input their interests to get customized stories tailored to their passionate engagement. Apart from commercial applications it can have commercial applications in various industries and it can automate the generation of promotional content and advertisements to Bengali speaking audience. Overall, the generation of Bengali stories has various practical implications across cultural, entertainment and etc. to enrich Bengali literature and Bengali language research.

## 5.3 Future Works

- We will improve our model to properly understand user input with advances in natural language understanding (NLU).
- We will increase size of data stories(first target 10000) to improve the quality with generate more coherent sentences of our model.
- We will also work with text to voice translation for bengali where we will be able to listen to the generated stories.
- For insufficient data stories, the model can not provide much good output and we will apply another new approach over the current dataset to analyse performance.
- We will work on narrative or speech type story generation in Bengali where narrate will be generate text for main character of story.

# References


[1] M. Mouli, "Evolution of bangla," *The Daily Star*, Feb 21, 2019.

[2] A. Alabdulkarim, S. Li, and X. Peng, "Automatic story generation: Challenges and attempts," pp. 72–83, Jan 2021.

[3] "History," *Bangla Stories*.

[4] M. S. Islam, S. S. Sharmin Mousumi, S. Abujar, and S. A. Hossain, "Sequence-to-sequence bangla sentence generation with lstm recurrent neural networks," *Procedia Computer Science*, vol. 152, pp. 51–58, 2019. International Conference on Pervasive Computing Advances and Applications- PerCAA 2019.

[5] S. Abujar, A. K. M. Masum, S. M. M. H. Chowdhury, M. Hasan, and S. A. Hossain, "Bengali text generation using bi-directional rnn," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, 2019.

[6] M. R. Kibria and M. Yousuf, "Context-driven bengali text generation using conditional language model," *Statistics, Optimization Information Computing*, vol. 9, pp. 334–350, 03 2021.

[7] A. Alhussain and A. Azmi, "Automatic story generation: A survey of approaches," *ACM Computing Surveys*, vol. 54, pp. 1–38, 05 2021.

[8] N. McIntyre, "Learning to tell tales: Automatic story generation from corpora," *Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh*, vol. 65, pp. 1–208, 05 2011.

[9] M. P. Fay, "Driving story generation with learnable character models," *Massachusetts Institute of Technology*, vol. 93, pp. 1–111, 05 2014.

[10] H. Huang, C. Tang, T. Loakman, F. Guerin, and C. Lin, "Improving chinese story generation via awareness of syntactic dependencies and semantics," 2022.

[11] J.-W. Lin and R.-G. Chang, "Optimizing chinese story generation based on multi-channel word embedding and frequent pattern tree structure," 03 2021.

[12] J.-W. Lin, J.-H. Tseng, and R.-G. Chang, "Chinese story generation using conditional generative adversarial network," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 457–462, 2020.

[13] L.-G. Moreno-Jiménez, J.-M. Torres-Moreno, and R. S. Wedemann, "Automatic generation of literary sentences in french," *Algorithms*, vol. 16, no. 3, 2023.

[14] A. Hakami, R. Alqarni, M. Almutairi, and A. Alhothali, "Arabic poems generation using lstm, markov-lstm and pre-trained gpt-2 models," pp. 139–147, 09 2021.

[15] H. Hejazi, A. Khamees, M. Alshurideh, and S. Salloum, *Arabic Text Generation: Deep Learning for Poetry Synthesis*. 03 2021.

[16] S. García-Méndez, M. Fernández-Gavilanes, E. Costa-Montenegro, J. Juncal-Martínez, and F. Javier González-Castaño, "A library for automatic natural language generation of spanish texts," *Expert Systems with Applications*, vol. 120, pp. 372–386, 2019.

[17] G. Aguado, A. Bañón, J. Bateman, S. Bernardos, M. Fernández, A. Gómez-Pérez, E. Nieto, A. Olalla, R. Plaza, and A. Sánchez, "Ontogeneration: Reusing domain and linguistic ontologies for spanish text generation," in

*Workshop on Applications of Ontologies and Problem Solving Methods, ECAI*, vol. 98, 1998.

[18] J. Kumar, S. Shekhar, and R. Gupta, "Automatic headline generation for hindi news using fine-tuned large language models," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, p. 391–399, Dec. 2023.

[19] Ghude, Tejashree, Chauhan, Roshni, Dahake, Krushna, Bhosale, Atharv, and Ghorpade, Tushar, "N-gram models for text generation in hindi language," *ITM Web Conf.*, vol. 44, p. 03062, 2022.

[20] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," 2021.

[21] N. Peng, M. Ghazvininejad, J. May, and K. Knight, "Towards controllable story generation," in *Proceedings of the First Workshop on Storytelling* (M. Mitchell, T.-H. K. Huang, F. Ferraro, and I. Misra, eds.), (New Orleans, Louisiana), pp. 43–49, Association for Computational Linguistics, June 2018.

[22] A. Fan, M. Lewis, and Y. Dauphin, "Strategies for structuring story generation," 2019.

[23] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao, "Calliope: Automatic visual data story generation from a spreadsheet," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 453–463, 2021.

# List of Acronyms

| | |
|---|---|
| NLU | Natural Language Understanding |
| NLG | Natural Language Generation |
| LLMs | Large Language Models |
| SMT | Statistical Machine Translation |
| UTF-8 | Unicode Transformation Format-8 |
| RNN | Recurrent Neural Network |
| bridNG | Hybrid of Ngram and RNN(GRU) |
| GRU | Gated Recurrent Units |
| LSTM | Long Short Term Memory |
| NLP | Natural Language Processing |
| ML | Machine Learning |
| DL | Deep Learning |