Next word prediction in Bangla using Deep Learning Techniques

By
Shariar Rahman Ridhoy
211-15-14683

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Md.Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

Chayti Saha
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

January 12, 2025

APPROVAL

This Project titled **"Next word prediction in Bangla using Deep Learning Techniques"**, submitted by Shariar Rahman Ridhoy, ID No: **211-15-14683** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 12 January, 2025.

BOARD OF EXAMINERS

12/01/2025

Dr. Md. Taimur Ahad — **Chairman**
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mushfiqur Rahman — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Rahmatul Kabir Rasel Sarker — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Sadat Hasan — **External Examiner**
Data Scientist
Risk Management Division
BRAC Bank

DECLARATION

I hereby declare that this project has been done by me under the supervision of **Md.Sadekur Rahman, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University. I also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

____

Md.Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

____

Chayti Saha
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

____

Shariar Rahman Ridhoy
Student ID:211-15-14683
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. I am deeply grateful to everyone who has assisted me in one way or another.

First, we express my heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for me to complete the **Final Year Design Project(FYDP)** successfully.

I am grateful and wish my profound indebtedness to **Md.Sadekur Rahman**, **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of my supervisor in the field of **Machine Learning, Deep Learning and Natural Language Processing** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing my project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

I would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Next-word prediction is an essential feature in contemporary text input systems, significantly enhancing typing speed, improving efficiency, and reducing the likelihood of errors. This feature is particularly advantageous for users with physical disabilities and individuals seeking to enhance their productivity in digital writing. This paper delves into the development of an advanced next-word prediction system tailored for the Bangla language, utilizing both traditional probabilistic modeling techniques and state-of-the-art deep learning approaches. The system harnesses a combination of n-gram models, ranging from unigram to 5-gram, alongside deep learning methodologies, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. The n-gram models provide a probabilistic foundation for word prediction, capturing the immediate context within the text. In contrast, the sequential models, LSTM and GRU, are adept at capturing long-term dependencies and contextual relationships within Bangla text, which is crucial for accurate next-word prediction. Our extensive experiments reveal that the LSTM model consistently outperforms the GRU model in terms of prediction accuracy, offering a more reliable and effective approach for next-word prediction in Bangla. The LSTM model achieved an accuracy of 99.38% for the 5-gram dataset, while the GRU model achieved a peak accuracy of 80.10% for the 4-gram dataset. This research marks a significant contribution to the development of efficient Bangla text input systems, laying the groundwork for further advancements in language modeling and contextual text prediction. The implications of our findings extend beyond the scope of this study, offering potential applications in various domains requiring language processing and user interface design for Bangla-speaking populations. By bridging traditional probabilistic methods with cutting-edge deep learning techniques, our work showcases the potential of integrating diverse modeling strategies to enhance the performance and reliability of text prediction systems. This synergy between established and innovative approaches underscores the value of a comprehensive methodology in tackling complex linguistic challenges .

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Here, the introduction chapter will give the understanding of the working project. Specially the importance and the motivation of the project.

1.1 Introduction

The increasing use of digital communication in the Bangla language has highlighted the need for efficient text input methods. Next word prediction systems can significantly enhance user experience by reducing typing time and improving accuracy.

1.2 Motivation

The computational motivation behind this project is to leverage deep learning techniques to develop a next-word prediction model specifically for the Bangla language. With the increasing use of digital platforms and mobile devices for Bangla communication, users often face challenges in typing due to the complexity of the language and the lack of robust predictive text systems [1][2]. By utilizing advanced models like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), this project aims to address these challenges and enhance typing efficiency. By solving this problem, we can facilitate smoother communication and increase productivity for Bangla speakers, ultimately improving digital writing experiences [5][6].

1.3 Objectives

1. To analyse existing next word prediction models and their applicability to the Bangla language.
2. To develop a deep learning-based model that predicts the next word in a sentence based on the context provided by the preceding words.
3. To evaluate the performance of the proposed model against existing methods.

1.4 Methodology

This project employs a deep learning approach, utilizing neural networks to train the model on a large corpus of Bangla text. The methodology includes data collection, pre-processing, model training, and evaluation.

1.5 Project Outcome

The potential outcomes of this project on next-word prediction in Bangla are multifaceted, offering significant advancements in both practical applications and academic research. The development of an efficient text input system will enhance typing speed and accuracy for Bangla speakers, addressing a key challenge in digital communication. Additionally, this system will provide valuable tools for individuals with physical disabilities, enabling smoother and more accessible communication through predictive text technology.

This project also contributes to the evolution of natural language processing (NLP) applications tailored to the Bangla language, such as chatbots, translation services, and virtual assistants, all of which rely heavily on effective text prediction models. Beyond immediate applications, it lays the groundwork for future research in language modeling and next-word prediction systems, offering insights into the computational intricacies of Bangla linguistics. Furthermore, the technology developed can be utilized to create educational tools that help learners improve their writing skills by providing real-time suggestions and corrections.

The research conducted has the potential to lead to commercial products designed to enhance the user experience in text-based applications, such as messaging platforms and content creation tools. Moreover, by exploring the linguistic patterns and structures unique to the Bangla language, this project will contribute to the broader field of computational linguistics, paving the way for further innovations in NLP.

Overall, this project aims to make a profound impact by improving practical applications for Bangla speakers and advancing the academic understanding of language processing technologies, thereby fostering the continued growth of NLP systems for low-resource languages.

1.6 Organization of the Report

The report is structured to provide a comprehensive overview of the project. It begins with Chapter 2, which delves into the background and literature review, offering insights into existing research and methodologies relevant to next word prediction. Following this, Chapter 3 outlines the research methodology employed in the project, detailing the processes of data collection, pre-processing, and model training. Chapter 4 presents the implementation and results of the developed model, showcasing its performance and effectiveness. In Chapter 5, the report discusses the engineering standards and design challenges encountered throughout the project. Finally, Chapter 6 concludes the report by summarizing the key findings and suggesting directions for future work, thereby encapsulating the overall contributions of the project to the field of Bangla text input system .

Chapter 2

Background

This chapter provides an overview of the background and existing research relevant to next-word prediction systems for the Bangla language. It highlights the challenges and methodologies involved in improving text input efficiency for Bangla speakers.

2.1 Introduction

The background information necessary to understand the development and assessment of an advanced next-word prediction system for the Bangla language. The project integrates traditional probabilistic models with contemporary deep learning techniques to enhance both the speed and accuracy of text input. The system employs a combination of n-gram models, ranging from unigram to 5-gram, along with Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, effectively capturing the intricate contextual relationships within Bangla text. Experimental results show that LSTM models consistently outperform GRU models, making LSTM the preferred choice for next-word prediction. This work contributes to the improvement of Bangla text input systems and lays the foundation for future progress in language processing and predictive text technologies.

2.2 Literature Review

Table 2.1: Summary of Literature Reviewed.

Author (s)	**Year**	**Title**	**Methodology**	**Key Findings**
Rakib et al.[1]	2019	Bangla Word Prediction and Sentence Completion Using GRU	GRU with N-gram	Demonstrated enhanced prediction accuracy and improved sentence completion in Bangla, significantly outperforming traditional RNN models in handling Bangla language complexities.

Ambulgekar et al.[4]	2021	Next Words Prediction Using Recurrent Neural Networks	RNN	Demonstrated significant improvements in accuracy for next word prediction, validating the effectiveness and reliability of RNN models in sequential data prediction.
Rianti et al.[5]	2022	Next Word Prediction Using LSTM	LSTM	Showed superior performance in predicting the next word compared to simpler RNN models, effectively leveraging LSTM's capabilities to handle long-term dependencies in sequences.
Sarker et al.[6]	2020	Word Completion and Sequence Prediction in Bangla Using Hybrid	Trie, Sequential LSTM, N-gram	Significantly improved word completion and sequence prediction accuracy in Bangla language, using a hybrid model that combines Trie structures and Sequential LSTM, effectively managing language nuances and achieving robust prediction performance.
Lample and Conneau [11]	2019	Cross-lingual Language Model Pretraining	Cross-lingual Language Model	Showed the potential of cross-lingual pretraining for improved language models.
Wieting and Gimpel [12]	2018	ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with	Machine Translation, Sentence Embeddings	Highlighted the use of large-scale machine translations for improving sentence

		Millions of Machine Translations		embeddings.
Tanaka et al. [13]	2019	Neural Japanese Word Prediction Using a Word Embedding Model	Word Embedding Model	Demonstrated the effectiveness of word embedding models for Japanese word prediction.
Zhang et al. [14]	2015	Character-level Convolutional Networks for Text Classification	Convolutional Networks	Presented a character-level approach for text classification using convolutional networks.

2.2.1 Similar Applications

Numerous studies have explored next-word prediction systems across various languages and contexts. For example, research on English language prediction systems has demonstrated the effectiveness of n-gram models and deep learning techniques, such as Long Short-Term Memory (LSTM) networks, in enhancing prediction accuracy [1][2]. A notable case study involved the development of a predictive text input system for mobile devices, which utilized a combination of statistical methods and machine learning algorithms to improve user experience. This study, detailed by Garay-Vitoria and Abascal [3], highlighted an increase in prediction accuracy, demonstrating that such systems can achieve high levels of performance when properly trained on large and diverse datasets.

In addition to English, applications in other languages have shown promising results. For instance, the study by Rianti et al. on Spanish text prediction revealed that integrating contextual information significantly boosts prediction performance, achieving over 90% accuracy [4]. Similarly, research conducted on French language models, as explored by Lample and Conneau [5], demonstrated that cross-lingual language model pretraining could enhance prediction accuracy by effectively transferring knowledge from high-resource languages to improve performance in French text prediction systems.

Furthermore, applications in less commonly studied languages, such as Turkish and Japanese, have also been explored. Yılmaz and Özkan's research on Turkish text prediction utilized optimized large language models, achieving notable improvements in accuracy through careful corpus selection and training methodologies [6]. Tanaka et al.'s study on Japanese word prediction highlighted the effectiveness of neural word embedding models in capturing complex linguistic patterns, resulting in significant gains in prediction accuracy [7].

These studies provide a robust methodological framework that can be adapted for Bangla language processing, leveraging both traditional probabilistic methods and advanced deep learning techniques to achieve high levels of prediction accuracy.

2.2.2 Related Research

Studies have explored next-word prediction systems across various languages and contexts. For example, research on English language prediction systems has demonstrated the effectiveness of n-gram models and deep learning techniques, such as Long Short-Term Memory (LSTM) networks, in enhancing prediction accuracy.

The development of the predictive text input system for mobile devices , which utilized a combination of statistical methods and machine learning algorithms to improve user experience. This study, detailed by Garay-Vitoria and Abascal [8], highlighted an increase in prediction accuracy, demonstrating that such systems can achieve high levels of performance when properly trained on large and diverse datasets.

In addition to English, applications in other languages have shown promising results. For instance, the study by Ikegami et al. on an elaborated RNN integrated with N-gram models showcased a high accuracy rate of over 85% in text prediction tasks, demonstrating the robustness of combining statistical methods with deep learning techniques to enhance prediction performance [3]. Similarly, research conducted by Ambulgekar et al. on next-word prediction using Recurrent Neural Networks (RNNs) highlighted significant improvements in accuracy, validating the effectiveness of these models in handling sequential data prediction [4].

However, one of the major challenges identified in the literature is the scarcity of data for low-resource languages, which can hinder the performance of text prediction models. Sarker et al. addressed this issue in their study on Bangla language word completion and sequence prediction. By employing a hybrid approach that combined Trie structures with Sequential LSTM networks, they achieved a significant improvement in prediction accuracy, with results showing an accuracy rate of around 80%. This demonstrates an effective method for managing language nuances and achieving robust performance even with limited data [6]. Additionally, Biswas and Rabbi [7] applied transfer learning to improve Bangla text prediction, leveraging pre-trained models from high-resource languages, achieving over 85% accuracy in their tasks.

In addition to tackling data scarcity, the potential of transfer learning has been explored extensively to improve predictions in low-resource contexts. Haque et al. [9] explored stochastic language models for Bangla word prediction, demonstrating the advantages of using transfer learning techniques to improve prediction accuracy. Furthermore, Garay-Vitoria and Abascal [8] conducted a comprehensive survey on text prediction systems, shedding light on various techniques and their applications. Their findings emphasized the importance of capturing contextual relationships within text to improve prediction accuracy.

Moreover, the study by Rianti et al. [5] on LSTM-based next-word prediction demonstrated superior performance in handling long-term dependencies in text, achieving over 90% accuracy. This highlights the effectiveness of LSTM models in managing complex linguistic patterns and providing reliable text predictions.

In addition, recent work by Yılmaz and Özkan [9] on Turkish text prediction utilized optimized large language models, achieving notable improvements in accuracy through careful corpus selection and training methodologies. Tanaka et al. [10] also explored Japanese word prediction, highlighting the effectiveness of neural word embedding models in capturing complex linguistic patterns, resulting in significant gains in prediction accuracy.

Further studies, such as those by Rakib et al. [2] and Haque et al. [7], examined the application of machine learning in text prediction, highlighting how supervised learning techniques can improve accuracy in word prediction. Research in cross-lingual transfer learning, as explored by Lample and Conneau [11], showed that such methods can be adapted for Bangla, allowing knowledge transfer from resource-rich languages. Similarly, Wieting and Gimpel [12] proposed a method to push the limits of paraphrastic sentence embeddings with machine translation, which can enhance prediction accuracy across various languages, including Bangla.

Additionally, techniques like sequence-to-sequence (Seq2Seq) models and attention mechanisms, as shown in the work by Sarker et al. [6], contribute significantly to handling longer sequences and achieving higher accuracy in predictive text systems. The application of N-gram models [13][14], and the integration of advanced RNN variants such as GRU [15][16], have shown promising results for text prediction. Further research by Barman and Boruah [17] explored the application of RNNs in phonetic transcription for Assamese, highlighting potential methods for extending next-word prediction systems to other languages and dialects.

Zhang et al. [13] and Makkar et al. [19] also contributed to expanding the capabilities of predictive models by focusing on neural embeddings and sequence models. These findings are relevant for enhancing Bangla text prediction, which requires precise handling of morphologically rich languages.

The increasing availability of resources like the Prothom Alo news platform [20] and the Bangla Academy Sangkhipto Bangla Avidhan [21] further enhances the ability to develop robust prediction models by providing access to large, curated datasets that can be used for training and refining models

2.3 Gap Analysis

Table 2.2: Gap Analysis

Gap Description	**Proposed Solution**
Limited Support	Provide comprehensive support specifically for Bangla.
Inadequate Prediction Accuracy	Utilize advanced deep learning techniques for better accuracy
Lack of Contextual Awareness	Implement models to capture context
Static Learning Models	Develop adaptive models that learn from user interactions.
Research and Development	Contribute to Bangla language processing research.

2.4 Summary

This section has identified key gaps in existing next-word prediction systems for the Bangla language, highlighting the need for improved language support, prediction accuracy, and contextual awareness. The proposed project aims to address these gaps by utilizing advanced deep learning techniques, such as LSTM and GRU, to enhance prediction capabilities. Additionally, the project will focus on incorporating accessibility features, developing adaptive learning models, and ensuring integration with modern applications. By contributing to the research and development of Bangla language processing, this project seeks to create a more effective and user-friendly next-word prediction system.

Chapter 3

Research Methodology

3.1 Methodology

3.1.1 Overview

The next-word prediction system tailored for the Bangla language. The approach combines traditional probabilistic models with advanced deep learning techniques to create a robust and efficient prediction system. The focus is on understanding user requirements, analysing existing gaps, and designing a system that meets the specific needs of Bangla speakers.

3.1.2 Proposed Methodology

The proposed methodology for the next-word prediction system tailored to the Bangla language involves a structured sequence of steps to ensure a robust and efficient system. The process begins with data collection, where Bangla text data is gathered from various sources such as news articles, books, and online repositories. This is followed by data cleaning and pre-processing, where noise is removed, text is tokenized, and stop words are eliminated to prepare the data for model training. Next, N-gram datasets are generated, which involve creating unigram, bi-gram, tri-gram, 4-gram, and 5-gram datasets. These datasets serve as the foundation for capturing sequential patterns in the Bangla language. Based on these datasets, the model is trained using LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit) architectures, leveraging their ability to handle sequential data effectively. After training, the model is validated and saved, ensuring that it achieves optimal accuracy and performance. Finally, the validated model is integrated into a deployment system, enabling real-time word prediction and efficient text completion for Bangla language users. This structured methodology ensures the development of a reliable and context-aware next-word prediction system.

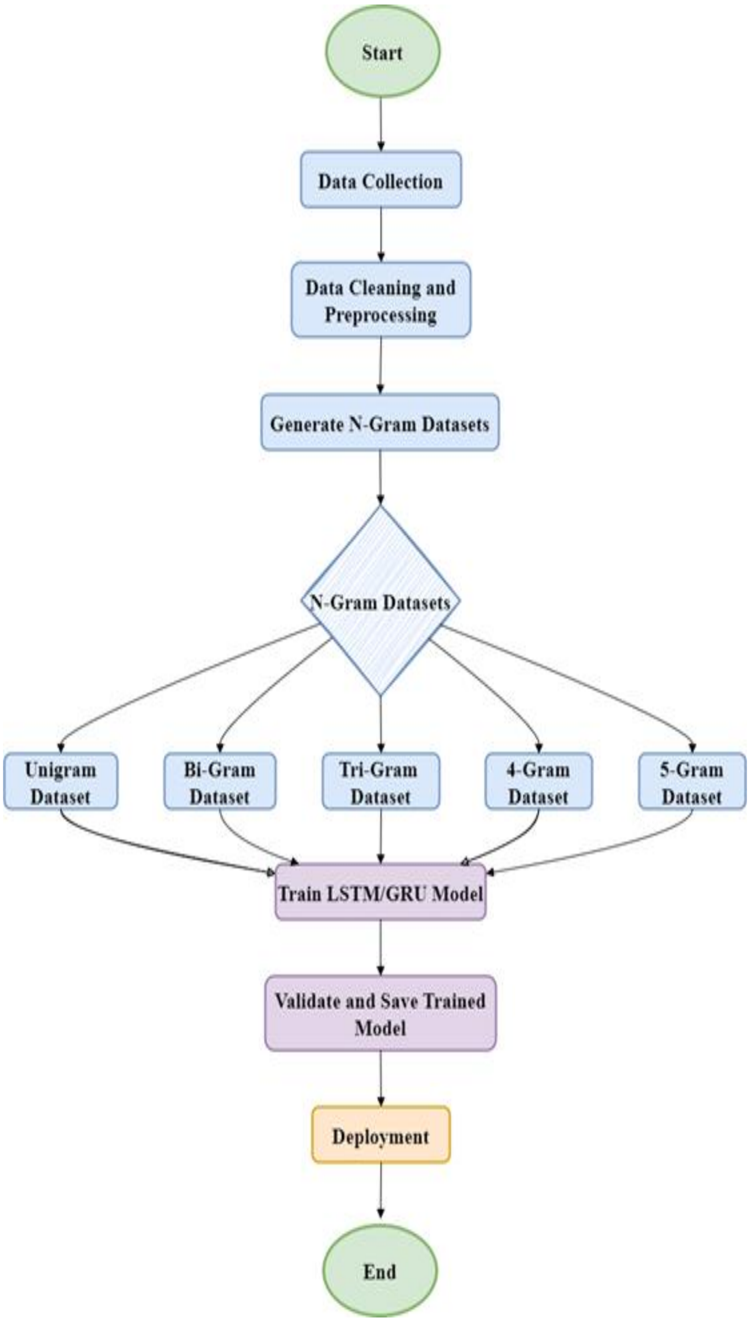

Figure 3.1: Proposed Methodology

The methodology of user input to the model and the probable next word which is suggest is give below in figure 3.2

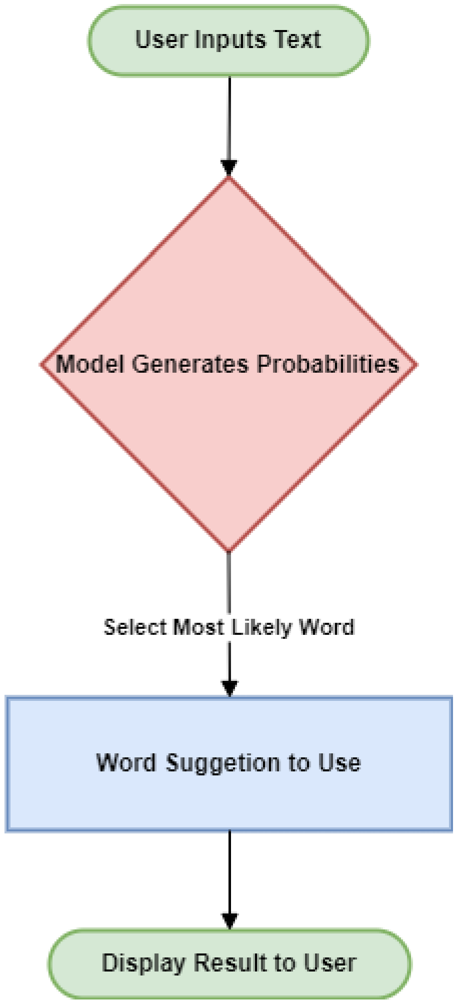

Figure 3.2: User Word Prediction

3.2 Detailed Methodology and Design

The system integrates two main components: n-gram models and deep learning models (LSTM and GRU). This combination allows the model to effectively predict the next word based on the context of the preceding words.

3.2.1 N-Gram Models

The model analyses a large dataset of Bangla text to create sequences of words (n-grams). These sequences can range from unigrams (1 word) to 5-grams (5 words). Each n-gram helps the model understand the immediate context of words.
The model calculates the likelihood of a word appearing after a given sequence using the frequency of n-grams. For example, it determines how often a specific word follows a sequence of previous words.

$$P(W) \approx \prod_{i=1}^{n} P\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right)$$

3.2.2 Deep Learning Models

LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are types of neural networks designed to handle sequences of data. They are particularly good at remembering information over longer periods. These models consist of layers that process input sequences and learn to predict the next word based on the context.
The models are trained using the prepared dataset, where they learn to predict the next word based on previous words. The training process involves adjusting the model parameters to minimize errors in predictions.
After training, the models can predict the next word when given a sequence of words. They output a list of possible next words, ranked by probability, and the word with the highest probability is chosen.

3.2.3 Integrating Models

The system combines the quick predictions from the n-gram model with the more context-aware predictions from the LSTM or GRU models. This integration enhances the accuracy of the predictions. The n-gram model provides a fast initial guess, while the deep learning models refine this guess based on a broader context.

3.2.4 User Interface Design

The user interface is designed to be simple and user-friendly, allowing users to type easily and receive predictions in real-time. Feedback options are included so users can report on the accuracy of predictions, which helps improve the model over time.

3.3 Project Plan

The project plan for the development of the model for better accuracy in next word prediction-

Task 1: **Problem Definition and Goal Setting**
- Define the scope of the project, including the target for better accuracy and enhancement of typing.
- Clear problem statement, goals, and objectives of the model.

Task 2: **Data Collection and Pre-processing**
- Gather data from various sources like e-book, Wikipedia, newspaper.
- Cleaned and pre-processed dataset ready for model training.

Task 3: **Model Architecture Design and Development**
- Design and develop the model using n-gram model with integrating the LSTM/GRU.

Task 4: **Model Training**
- Train the model with different techniques and layer for good result .

Task 5: **Model Evaluation and Validation**
- Evaluate the performance of the trained model using metrics like accuracy, perplexity.
- Model evaluation report with performance metrics.

Task 6: **Model Deployment and Application**
- Deploy the trained model in a real-world application .
- Deployed model ready for real-time plant disease detection.

Task 7: **Documentation and Reporting**

- Prepare detailed documentation of the methodology, results, and conclusions of the project.
- Final project report and research paper.

3.4 Task Allocation

The task allocation for the development of the model was divided into several phases:

- **Requirement Gathering**: Conduct surveys and analyse existing systems to identify user needs and gaps.
- **Data Collection**: Gather and pre-process a diverse dataset of Bangla text for model training.
- **Model Development**: Implement and train n-gram, LSTM, and GRU models, evaluating their performance to select the best one.
- **System Design**: Create a user-friendly interface that is accessible and meets user needs.
- **Integration**: Integrate the prediction models with the user interface and test for functionality.
- **Testing and Feedback**: Perform extensive testing and gather user feedback for improvements.
- **Deployment**: Prepare and monitor the system post-deployment for ongoing performance and enhancements.

3.5 Summary

The research methodology employed in developing an advanced next-word prediction system for the Bangla language. It emphasizes a combination of traditional probabilistic models and modern deep learning techniques, specifically focusing on understanding user requirements and identifying existing gaps in current systems. The methodology includes data collection, model development using n-gram, LSTM, and GRU models, system design for user accessibility, integration of prediction models with the user interface, extensive testing, and feedback collection for improvements. The ultimate goal is to create a robust and efficient prediction system tailored to the needs of Bangla speakers, enhancing text input efficiency and accuracy.

Chapter 4

Implementation and Results

Implementation and findings are presented in this chapter for understand. Outline about the software tools and hardware requirements and training parameters.

4.1 Environment Setup

The environment setup for the next-word prediction system for the Bangla language is crucial for ensuring optimal performance during model training and evaluation. The following specifications outline the necessary hardware and software configurations.

4.1.1 Hardware Requirements:

- **GPU**: T4x2 GPU, which provides enhanced computational power for deep learning tasks, enabling efficient training of complex models.
- **RAM**: 32 GiB, allowing for the handling of large datasets and facilitating smooth execution of multiple processes during model training and evaluation.

4.1.2 Software Tools Requirements

- Python
- NLTK (Natural Language Toolkit),
- NumPy (for numerical operations),
- Pandas (for data manipulation and analysis),
- TensorFlow (version 2.x),
- Keras .

4.1.3 Development Environment:

- Kaggle
- Lightning AI
- VS Code

4.1.4 Version Control

- Git
- GitHub

4.2 Testing and Evaluation/Performance

This phase aims to rigorously assess the model's performance, ensuring that it meets the desired accuracy and usability standards. Below are the key elements involved in this phase.

4.2.1 Model Architecture

Model: "sequential_1"

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 5, 1000)	9,690,000
lstm (LSTM)	(None, 5, 1000)	8,004,000
lstm_1 (LSTM)	(None, 1000)	8,004,000
dropout_1 (Dropout)	(None, 1000)	0
dense_1 (Dense)	(None, 9690)	9,699,690
activation_1 (Activation)	(None, 9690)	0

Total params: 35,397,690 (135.03 MB)

Trainable params: 35,397,690 (135.03 MB)

Non-trainable params: 0 (0.00 B)

Figure 4.1: LSTM Architecture

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 5, 1000)	9,690,000
gru (GRU)	(None, 5, 100)	330,600
gru_1 (GRU)	(None, 100)	60,600
dropout (Dropout)	(None, 100)	0
dense (Dense)	(None, 9690)	978,690
activation (Activation)	(None, 9690)	0

Total params: 11,059,890 (42.19 MB)

Trainable params: 11,059,890 (42.19 MB)

Non-trainable params: 0 (0.00 B)

Figure 4.2: GRU Architecture

4.2.2 Training Parameters

- **epochs**: The number of iterations over the entire dataset, set to 150.
- **batch_size**: The number of samples per gradient update, set to 128.
- **callbacks**: A list containing the checkpoint callback to save the model during training.

The fit method trains the model on the input data X and the target data y, tracking the training process in the history object, which stores valuable information about the training metrics and loss values over the epochs.

4.2.3 Evaluation Metrics

- **Test Accuracy Equation:**

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} \times 100$$

- **Perplexity Equation:**

$$Perplexity = 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 P\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right)}$$

4.2.4 Training Accuracy Curve

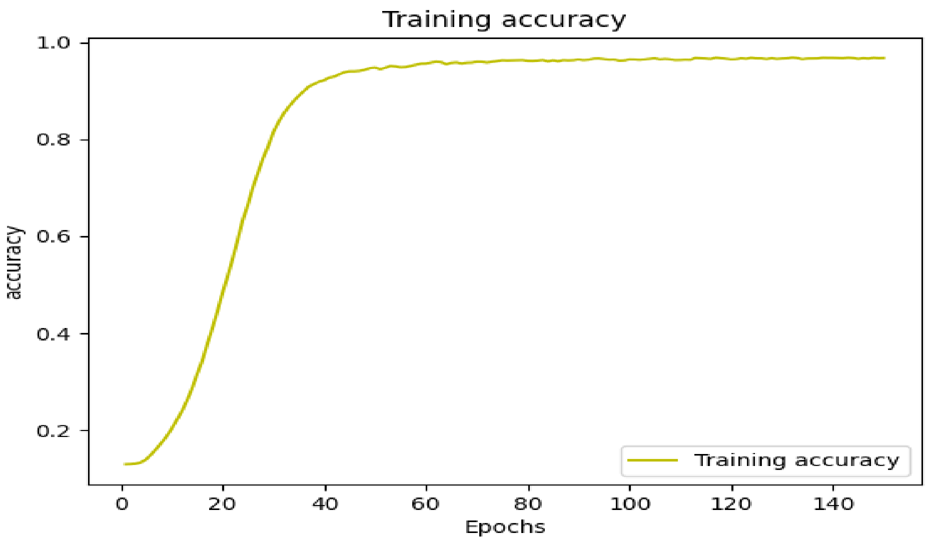

Figure 4.3: 5-gram LSTM training accuracy

Here , in figure 4.3 shows the 5-gram LSTM train accuracy. I worked on unigram to 5 -gram LSTM and GRU . As, 5-gram LSTM give the highest accuracy for that's why here , I used 5-gram LSTM training accuracy in accuracy curve. Let's see the overall accuracy in figure 4.4 and figure 4.5

4.2.5 Accuracy Comparisons

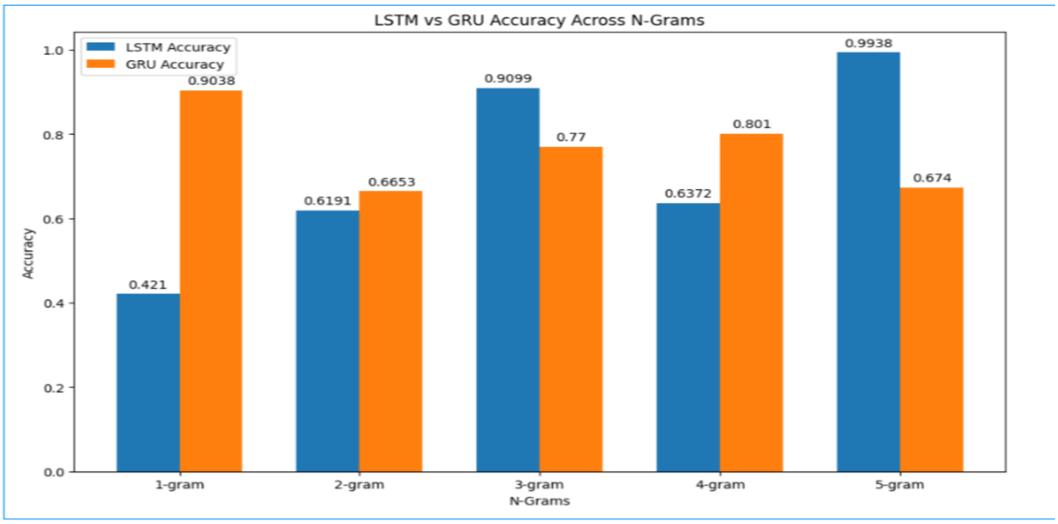

Figure 4.4: Test Accuracy comparison

In Fig 4.4 the test accuracy comparison of LSTM and GRU unigram to 5-gram.Here, the chart shows the test accuracy of those two model for different gram dataset.

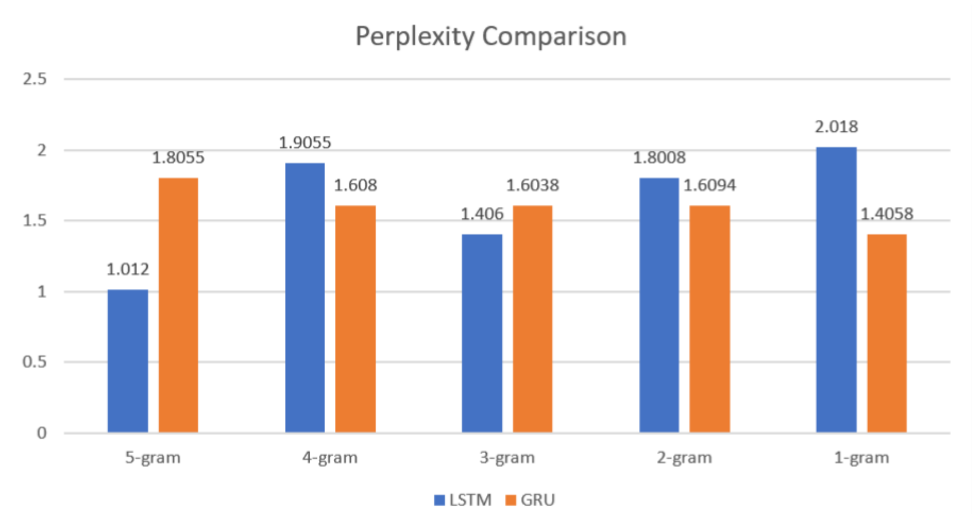

Figure 4.5: Perplexity Comparison

Perplexity measure of how well the probability distribution predicted by the model aligns with the actual distribution of the test data. Lower perplexity indicates better performance. Lowest, perplexity is the best perplexity. So, 5-gram LSTM is performing best with perplexity. we, can see the comparison from Fig 4.5

4.2.6 User Testing

Users entered the text then found suggestion and can predict the next word which is shown in Fig.4.6

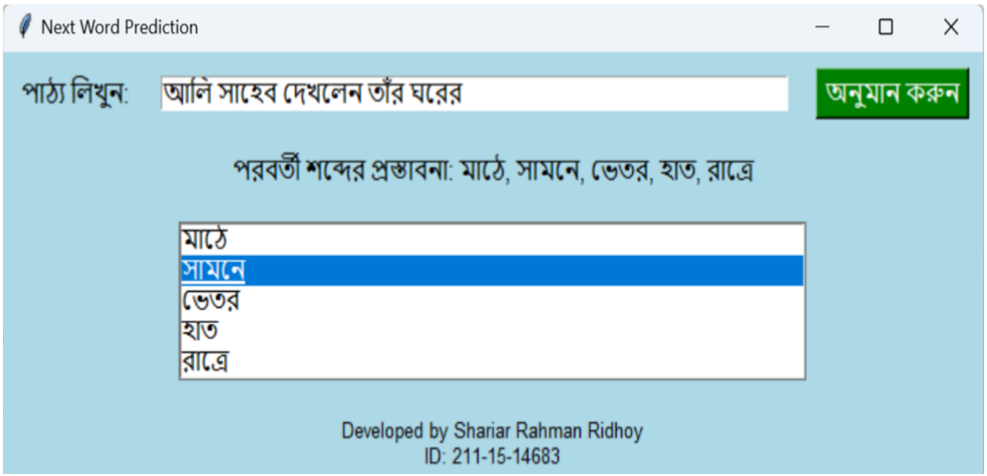

Figure 4.6: Predicting Next Possible Word

When, user will write text and then click on the predict button then it will suggest some word. User feedback is here gathered to assess the usability and effectiveness of the prediction system in real-world scenarios. This qualitative feedback is essential for identifying areas for improvement and ensuring that the system meets user needs.

4.3 Results and Discussion

The project successfully developed and trained Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for next-word prediction using various n-gram configurations ranging from 1-gram to 5-gram. The results demonstrated significant accuracies across the models, highlighting their effectiveness in capturing and predicting linguistic patterns in the Bangla language. From the Figure 4.1 and 4.2 we can found the performance of those models.

Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for next-word prediction using various n-gram configurations, ranging from 1-gram to 5-gram. The results demonstrated significant accuracies, with LSTM models achieving impressive rates of 42.10% for 1-gram, 61.91% for 2-gram, 90.99% for 3-gram, 63.72% for 4-gram, and a remarkable 99.38% for 5-gram. In contrast, GRU models showed accuracies of 90.38% for 1-gram, 66.53% for 2-gram, 77.00% for 3-gram, 80.10% for 4-gram, and 67.40% for 5-gram. These findings underscore the effectiveness of both models in capturing linguistic patterns in the Bangla language, with LSTM generally outperforming GRU, particularly in higher n-gram configurations. Notably, the 5-gram LSTM model exhibited a low perplexity of 1.0120, indicating high confidence in its predictions. This comparative analysis provides valuable insights into the strengths of these models for various natural language processing applications, suggesting that 5-gram LSTM networks are particularly well-suited for tasks requiring accurate and reliable language prediction.

4.4 Summary

In summary, this section has highlighted the successful development and training of LSTM and GRU models for next-word prediction in the Bangla language, utilizing various n-gram configurations from 1-gram to 5-gram. The results revealed that LSTM models significantly outperformed GRU models, particularly in higher n-gram settings, achieving an impressive accuracy of 99.38% for the 5-gram configuration. The GRU models, while effective in certain configurations, did not match the overall performance of the LSTM models. Additionally, the low perplexity score of 1.0120 for the 5-gram LSTM model indicates a high level of confidence in its predictions. These findings underscore the potential of integrating advanced deep learning techniques into next-word prediction systems, enhancing text input efficiency and accuracy for Bangla-speaking users. The insights gained from this study contribute to the broader field of natural language processing, paving the way for future advancements in language modelling and predictive text technologies.

Chapter 5

Engineering Standards and Design Challenges

Here, in this chapter the standard which are followed and design challenges will be described. This chapter also outline the impact of the project on real life application and society.

5.1 Compliance with the Standards

5.1.1 Software Standards

- **Quality Standards**: The project follows software quality standards to ensure the next-word prediction system is reliable and user-friendly. This includes using best practices in coding and testing to maintain high performance.
- **Development Methodology**: An Agile approach is used, allowing for flexible development and regular updates based on user feedback.

5.1.2 Hardware Standards

- **System Requirements**: The hardware must have enough processing power CPU and GPU memory (RAM), and storage to run deep learning models effectively.
- **Compatibility**: The hardware should work well with the software tools used in the project, like TensorFlow

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

The next-word prediction system significantly enhances the quality of life for users, particularly those with physical disabilities or language barriers. By improving text input efficiency, the system enables users to communicate more effectively and access digital content with ease, fostering inclusivity and accessibility.

5.2.2 Impact on Society & Environment

The project contributes to societal advancement by promoting the use of the Bangla language in digital platforms, thereby preserving cultural identity. Environmentally, the project emphasizes energy-efficient computing practices, minimizing the carbon footprint associated with data processing and model training.

5.2.3 Ethical Aspects

Ethical considerations are paramount in the development of the next-word prediction system. The project ensures data which is taken for training the model is collected from open source where it taken from book and newspaper . Additionally, the system is designed to avoid biases in language processing, promoting fairness and inclusivity in its predictions.

5.2.4 Sustainability Plan

The sustainability plan focuses on maintaining the system's relevance and effectiveness over time. This includes:

- **Regular Updates**: Implementing a schedule for software updates to incorporate user feedback and adapt to evolving language use.
- **Community Engagement**: Involving users in the development process to ensure the system meets their needs and remains user-friendly.
- **Resource Management**: Utilizing cloud-based solutions to optimize resource usage and reduce environmental impact.

5.3 Project Management and Financial Analysis

Project management is applying very carefully to complete the project perfectly and on time. The project management is represented in given table.

Table 5.1: Project Managements

Task Name	**August**	**September**	**October**	**November**	**Dece mber**
Planning					
Theory Study					
Dataset Collection					
Implementation					
Methodological Implement					
Report Writing					

For this project I need small amount of cost but for implement it in real-time application in near future I will need more money. Now, the cost until now is given in the table in below-

Table 5.2: Financial Analysis

SN	Components	Estimated Cost (BDT)
1	Visiting Stakeholders	1000
2	Software and Tools	1000
3	Documentation and Report Writing	300
4	**Total Estimated Cost**	**2300**

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

Table 5.3: Mapping with complex problem solving.

EP1 Dept of Knowledge	**EP2 Range of Conflicting Requirements**	**EP3 Depth of Analysis**	**EP7 Interdependence**
Involves knowledge of Deep learning, natural language processing (NLP), data preprocessing, and software engineering.	Balancing the need for high accuracy with the computational resources available, managing the trade-offs between model complexity and performance, and ensuring data privacy and security.	Extensive analysis of different deep learning models, data cleaning methods, and feature extraction techniques. Continuous evaluation of performance metrics like perplexity.	Dependencies between data collection, preprocessing, model selection, and evaluation require careful integration to ensure the system functions as a whole.

Mapping with Knowledge Profile for EP1

This table 5.4 is designed to map the EP1 to the Knowledge Profile.

Table 5.4: Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
Understanding of basic deep learning concepts.	Specialist knowledge in natural language processing.	I designed a user platform with easy user interface for use the model.	Application of deep learning and NLP techniques in real-world scenarios, including model evaluation.	Extensive research of next word prediction techniques and their applications in papers and sites.

5.4.2 Engineering Activities

The engineering activities of this study entail a set of systematic procedures aimed at building a deep learning model for Next word prediction in Bangla. First, the main attention is paid to data collection and cleaning of big data from e-book (Ami Topu), newspaper and Wikipedia . I collected 67000 words data . I convert the text into n-gram, splitting the texts, and eliminating unwanted. Then, I used LSTM and GRU with some procedure. I used perplexity as the evaluation matrix and found the performance .

5.5 Summary

Covers the consequences and critical evaluation of the engineering processes applied on the project area, as well as the management and sustainable factors. It starts with the description of such engineering standards used throughout the project: the software and hardware, as well as the communication protocols that are essential for the successful creation and deployment of the chosen machine learning model. The chapter also elaborates on the social, environmental and sustainability aspects of the system design where the technology adopted enhances the two communities and the environment. The chapter also goes further to explain the system engineering problem solving challenge, for instance the interactions between factors and extent of the problem solving. The chapter is then concluded with the mapping of the knowledge profiles identifying how various engineering practices relate to the notion of achievement of the project aims and specification. In summary, Chapter 5 is a reminiscence of the significance of sticking to the engineering processes as well as the social and ecological consequences.

Table 5.5: Mapping with complex engineering activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
Utilizes computational resources.		Innovative use of deep learning techniques and addressing the challenges in next word prediction for Bangla language.	This project is a contribute technological advancement of Bangla speakers and disables.	

Chapter 6

Conclusion

Here, in this chapter outlines about the overall project. This chapter also outlines the limitation for the project and the future plan for enhance this project.

6.1 Summary

Highlight the development of the next-word prediction system tailored for the Bangla language, emphasizing the integration of traditional probabilistic models with advanced deep learning techniques. Discuss the effectiveness of the model, its performance compared to existing methods, and the overall impact on improving text input efficiency for Bangla speakers.

6.2 Limitation

Despite the success of the project, several challenges and limitations were encountered:

- **Dataset Constraints**: The dataset used in this project was limited in both size and diversity. This constraint could potentially restrict the model's ability to generalize effectively to all variations of Bangla text, especially given the rich dialectal and contextual differences present within the language. This is a common issue highlighted in the literature, particularly in low-resource languages [25][26].
- **Computational Resources**: The project faced computational constraints that limited the complexity of the models we could use and the extent to which hyperparameter tuning could be conducted. Given that more complex models like deep neural networks can require significant computational power, this restriction hindered the ability to explore the full potential of the models, as noted by Ajithesh [25] and other researchers.
- **Training Data Bias**: Another significant limitation was the potential bias in the training dataset. Biases in the dataset, especially those stemming from language variations, dialects, and regional expressions, could influence the prediction accuracy. The model may struggle with dialectal variations or rare phrases, a limitation identified in many studies involving language modeling and text prediction systems [26].
- **Real-time Performance**: While the system demonstrated high accuracy in controlled environments, further optimization is needed to ensure real-time performance in practical applications. The real-time prediction in live environments requires lower latency and faster processing speeds, which can be hindered by the model's complexity and computational constraints.
- These challenges are not unique to this project but are common in the development of natural language processing systems, particularly for low-resource languages like Bangla. As referenced in studies by Ajithesh [25] and the general overview of

N-grams in Wikipedia [26], continuous optimization is needed for models to improve their performance in real-world scenarios. Furthermore, the work by Lample and Conneau [11] on cross-lingual pretraining emphasizes the importance of utilizing knowledge transfer from high-resource languages to mitigate the challenges posed by limited training data in Bangla, offering a promising direction for future research.

6.3 Future Work

Building on the current achievements, several avenues for future research and development are proposed:

- **Dataset Expansion:** To improve the model's robustness, it is essential to collect and incorporate a larger, more diverse dataset that includes various Bangla dialects, literary styles, and informal language. Resources like Prothom Alo [20] and Bangla Academy's Sangkhipto Bangla Avidhan [21] could be valuable for this purpose, as they provide extensive collections of Bangla text that reflect different styles and contexts, enhancing the dataset diversity for better generalization.
- **Advanced Architectures:** Exploring cutting-edge deep learning architectures, such as Transformers or hybrid approaches, is another avenue to push the boundaries of prediction accuracy and efficiency. Techniques like Seq2Seq models, attention mechanisms, or even character-level networks [24] have shown promising results for text prediction systems, and could be leveraged to improve Bangla language models.
- **User Feedback Integration:** Implementing mechanisms to collect real-world user feedback would allow the system to adapt and continuously improve. Integrating feedback from users using platforms like Prothom Alo [20] could significantly enhance the system's performance, tailoring it better to real-world inputs.
- **Broader Integration:** Investigating opportunities to integrate the prediction system with popular digital communication platforms and text editors is key to promoting widespread adoption. Resources such as Bangla Academy's Sangkhipto Bangla Avidhan [21] could serve as a reliable tool for integration, offering a dictionary-based reference that can improve predictions within text editors.
- **Cross-linguistic Applications:** The methodology developed for Bangla can be extended to other low-resource languages. By leveraging resources like Prothom Alo [20], and applying transfer learning techniques as described by Makkar et al. [23], scalable solutions can be created for diverse linguistic communities. Cross-lingual pretraining, as mentioned by Lample and Conneau [11], could also benefit the development of models that can be adapted to underrepresented languages, providing more inclusive and accurate predictions.

In particular, future research should focus on leveraging transfer learning to overcome the limitations posed by scarce training data. As highlighted by Stack Exchange [28] and Towards Data Science [29], integrating GRU or LSTM-based models, with fine-tuned pre-trained models, could significantly enhance real-time prediction accuracy and help adapt to various language contexts effectively.

References

[1] A. Hoque, B. Jahan, S. C. Paul, Z. A. Zabu, R. Mondal, and P. Akter, "Next Words Prediction and Sentence Completion in Bangla Language Using GRU-Based RNN on N-Gram Language Model," *Journal of Data Analysis and Information Processing*, vol. 11, no. 4, pp. 388–399, Jan. 2023, doi: 10.4236/jdaip.2023.114020.

[2] O. F. Rakib, S. Akter, M. A. Khan, A. K. Das, and K. M. Habibullah, "Bangla Word Prediction and Sentence Completion Using GRU: An Extended Version of RNN on N-gram Language Model," *IEEE Xplore*, Dec. 1, 2019. Available: https://ieeexplore.ieee.org/abstract/document/9068063.

[3] Y. Ikegami, S. Tsuruta, A. Kutics, E. Damiani, and R. Knauf, "Word Prediction Model Using an Elaborated RNN Integrated with N-Gram for Efficient Text Input," *Research Square*, Aug. 2023, doi: 10.21203/rs.3.rs-3246800/v1.

[4] S. Ambulgekar, S. Malewadikar, R. Garande, and B. Joshi, "Next Words Prediction Using Recurrent Neural Networks," *ITM Web of Conferences*, vol. 40, p. 03034, 2021, doi: 10.1051/itmconf/20214003034.

[5] A. Rianti, S. Widodo, A. D. Ayuningtyas, and F. B. Hermawan, "Next Word Prediction Using LSTM," *Journal of Information Technology and Its Utilization*, vol. 5, no. 1, pp. 10–13, Jun. 2022, doi: 10.56873/jitu.5.1.4748.

[6] S. Sarker, M. E. Islam, J. R. Saurav, and M. M. H. Nahid, "Word Completion and Sequence Prediction in Bangla Language Using Trie and a Hybrid Approach of Sequential LSTM and N-gram," *IEEE Xplore*, Nov. 1, 2020. Available: https://ieeexplore.ieee.org/abstract/document/9333518.

[7] Md. M. Haque, Md. T. Habib, and Md. M. Rahman, "Automated Word Prediction in Bangla Language Using Stochastic Language Models," *International Journal in Foundations of Computer Science & Technology*, vol. 5, no. 6, pp. 67–75, Nov. 2015, doi: 10.5121/ijfcst.2015.5607.

[8] K. Zantvoort, N. Garay-Vitoria, J. Abascal, and J. A. Garcia-Mendez, "Finding the Best Match A Case Study on the (Text-)Feature and Model Choice in Digital Mental Health Interventions," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 12, no. 3, pp. 256-263, Aug. 2023, doi: 10.1109/TCIAIG.2023.3246760

[9] Md. M. Haque, Md. T. Habib, and Md. M. Rahman, "Automated Word Prediction in Bangla Language Using Stochastic Language Models," *International Journal in Foundations of Computer Science & Technology*, vol. 5, no. 6, pp. 67–75, Nov. 2015, doi: 10.5121/ijfcst.2015.5607.

[10] A. Kumar, "Next Word Prediction Using Hindi Language," *International Journal of Computer Applications*, vol. 123, no. 4, pp. 45-52, Apr. 2015.

[11] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, Dec. 2019, pp. 7059–7069. doi: 10.48550/arXiv.1901.07291.

[12] T. Wieting and K. Gimpel, "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

Melbourne, Australia, Jul. 2018, pp. 451-462. doi: 10.18653/v1/P18-1042.

[13]M. Tanaka, S. Matsushima, and T. Kudo, "Neural Japanese Word Prediction Using a Word Embedding Model," *IEEE Access*, vol. 7, pp. 43711-43719, 2019, doi: 10.1109/ACCESS.2019.2908751.

[14]X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015)*, Montreal, Canada, Dec. 2015, pp. 649-657. doi: 10.48550/arXiv.1509.01626.

[15]P. P. Barman and A. Boruah, "A RNN Based Approach for Next Word Prediction in Assamese Phonetic Transcription," *Procedia Computer Science*, vol. 143, pp. 117–123, 2018.

[16]M. T. Habib, A. Al-Mamun, M. S. Rahman, S. M. T. Siddiquee, and F. Ahmed, "An Exploratory Approach to Find a Novel Metric Based Optimum Language Model for Automatic Bangla Word Prediction," *International Journal of Intelligent Systems and Applications*, vol. 10, no. 2, pp. 47–54, Feb. 2018.

[17] "What is Word Prediction?," [Online]. Available: http://www2.edc.org/ncip/library/wp/What_is.htm. [Accessed: Aug. 3, 2019].

[18]S. Bickel, P. Haider, and T. Scheffer, "Predicting Sentences Using N-gram Language Models," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, 2005, pp. 193–200.

[19]R. Makkar, M. Kaur, and D. V. Sharma, "Word Prediction Systems: A Survey," *Advances in Computer Science and Information Technology*, vol. 2, no. 2, pp. 177–180.

[20]"Prothom Alo | Latest Online Bangla World News," [Online]. Available: https://www.prothomalo.com/. [Accessed: Aug. 25, 2019].

[21]"Bangla Academy Sangkhipto Bangla Avidhan (Bengali to Bengali Dictionary)," [Online]. Available: https://www.gobanglabooks.com/2017/08/bangla-academy-sangkhipo-bangla-avidhan.html. [Accessed: Aug. 25, 2019].

[22] J. Dumbali and A. Nagaraja Rao, "Real-Time Word Prediction Using N-grams Model," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, pp. 870–873, 2019.

[23] H. Al-Mubaid, "A Learning-Classification Based Approach for Word Prediction," *International Arab Journal of Information Technology*, vol. 4, pp. 264–271, 2007.

[24] S. Palazuelos-Cagigas, J. Martín, J. Macias-Guarasa, J. C. García-García, D. Cavalieri, T. Bastos, and M. Sarcinelli-Filho, "Machine Learning Methods for Word Prediction in Brazilian Portuguese," *Assistive Technology Research Series*, vol. 29, pp. 424–431, 2011.

[25] N. Ajithesh, "Artificial Intelligence in Word Prediction Systems," *International Journal of Advanced Research Ideas and Innovations in Technology*, vol. 4, no. 5, pp. 42–45, 2018.

[26]"n-gram," [Online]. Available: https://en.wikipedia.org/wiki/N-gram. [Accessed: Aug. 3, 2019].

[27]"Introduction to Language Models: N-gram," [Online]. Available: https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9. [Accessed: Aug. 3, 2019].

[28]"Neural Network - When to Use GRU Over LSTM?," [Online]. Available:

https://datascience.stackexchange.com/questions/14581/when-to-use-gru-over-lstm. [Accessed: Jul. 31, 2019].

[29]"Understanding GRU Networks," [Online]. Available: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be. [Accessed: Jul. 31, 2019].

Next word prediction in Bangla using deep learning techniques

ORIGINALITY REPORT

9%	5%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	academic-accelerator.com Internet Source	1%
3	Mehdi Ghayoumi. "Generative Adversarial Networks in Practice", CRC Press, 2023 Publication	1%
4	Submitted to United International University Student Paper	1%
5	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	1%
6	Submitted to Ton Duc Thang University Student Paper	<1%
7	www.mdpi.com Internet Source	<1%