ARTIFICIAL INTELLIGENCE-BASED REGIONAL LANGUAGE ANALYSIS

BY

KABBO KUMAR
ID: 201-15-14293
AND

TANJIL HOSSAIN

ID: 201-15-13946

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

Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Ms. Israt Jahan
Lecturer (Sr. Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "**ARTIFICIAL INTELLIGENCE-BASED REGIONAL LANGUAGE ANALYSIS**", submitted by **Tanjil Hossain** and **Kabbo Kumar** 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 **13/07/2004**.

BOARD OF EXAMINERS

Hassain 13.07.2024

Dr. Md. Fokhray Hossain (MFH) — **Board Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Siddiquee

Mr. Shah Md. Tanvir Siddiquee (SMTS) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

13.7.24

Mr. Md Umaid Hasan (MUH) — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

13/7/24

Dr. Shamim H Ripon (DSR) — **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Shah Md Tanvir Siddiquee, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Ms. Israt Jahan
Lecturer (Sr. Scale)
Department of CSE
Daffodil International University

Submitted by:

TANJIL HOSSAIN
ID: 201-15-13946
Department of CSE
Daffodil International University

KABBO KUMAR
ID: 201-15-14293
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Mr. Shah Md Tanvir Siddiquee**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Natural language processing (NLP)" 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 them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori** (**Head of the Department of CSE)**,for his kind help in finishing our project and also to other faculty members and the staff ofthe Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Bangla, also known as Bengali, is a prominent language spoken by over 230 million people globally and serves as the official language of Bangladesh and one of the 22 scheduled languages of India. Despite its widespread usage, various local languages in Bangladesh, including Rajshahi, Chittagong, Sylhet, Barisal, and Rangpur, face endangerment due to urbanization, globalization, and limited educational opportunities. The shift towards dominant languages like Bengali, coupled with social stigma and limited exposure in media and formal settings, contributes to the gradual decline of local languages. To combat this decline, efforts are underway to preserve linguistic diversity and heritage. Artificial Intelligence (AI) and Machine Learning (ML) offer promising solutions to address the challenges local languages face. Language preservation tools powered by AI facilitate the documentation, transcription, and archiving of oral traditions, literature, and vocabulary in endangered languages. ML algorithms enable the creation of language learning platforms and mobile applications tailored to specific local languages, enhancing accessibility and engagement. Furthermore, AI-driven translation and localization technologies facilitate the integration of local languages into digital content and services, ensuring their relevance and accessibility in the digital age. Collaboration between AI researchers, linguists, educators, and community members is essential for the development and implementation of these solutions. Data collection for local languages can be achieved through direct engagement with local citizens, surveys, interviews, and community outreach programs, as well as from local literature and written materials. Leveraging AI and ML technologies holds promise in revitalizing and preserving local languages, thereby contributing to the preservation of linguistic diversity and cultural heritage in Bangladesh and beyond.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Overview

The vibrant linguistic landscape of Bangla encompasses a tapestry of dialects, each with unique nuances and complexities. However, traditional translation methods often struggle to navigate these subtle variations, hindering communication, understanding, and cultural preservation. This research project explores the potential of Artificial Intelligence (AI) to bridge these dialectal divides, focusing on translations between Dhakaia, Barishali, Sylheti, Rangpur, and Chittagong. This research goes beyond simply replicating existing models. It delves into the specific challenges posed by Bangla dialects, such as limited parallel data, dialect-specific vocabulary, and lack of standardized resources. To address these hurdles, we will explore techniques like data augmentation, domain adaptation, fine-tuning of model parameters, etc.

1.2 Background and Present State

The background of this project revolves around addressing the challenges in Bangla language processing, particularly in text sequence prediction. Bangla, as one of the world's major languages, lacks comprehensive natural language processing (NLP) tools and models compared to languages like English. Presently, there is a growing interest in developing robust NLP models tailored for Bangla to facilitate applications ranging from sentiment analysis to machine translation. The current state of Bangla NLP is marked by limited datasets, sparse linguistic resources, and a shortage of specialized tools. Existing models often face challenges in capturing the nuances of Bangla grammar, semantics, and context. Despite these hurdles, recent advancements in deep learning, particularly with recurrent neural networks (RNNs) and transformer models, show promise in improving the accuracy and applicability of NLP tasks for Bangla. This project aims to contribute to the advancement of Bangla NLP by exploring various deep learning architectures and methodologies to enhance text sequence prediction. By leveraging these advancements, we seek to address existing gaps and pave the way for more sophisticated applications of NLP in Bangla language processing.

1.3 Problem Statement

Bangladesh's linguistic landscape pulsates with diverse dialects, each carrying unique threads of cultural heritage. Yet, globalization and urbanization cast a shadow, threatening to silence the voices of lesser-known languages like Rajshahi, Chittagong, Sylhet, Barisal, and Rangpur. These dialects face a multitude of challenges:

- Limited Exposure**:** Their use in formal settings and media diminishes, reducing opportunities for younger generations to learn and engage.

- Standardization Gaps**:** Lack of standardized resources and documentation hinders their preservation and development.

- Technological Barriers**:** Existing translation methods struggle to capture the nuances and complexities of these dialects, hindering communication and cultural exchange.

1.4 Objectives

The potential of AI-powered translation to unlock seamless communication and understanding across Bangla dialects is deeply inspiring. This project is driven by a combined passion for both technological innovation and the preservation of linguistic diversity. By bridging the dialectal divide, we aim to foster a more connected and inclusive Bangla community where cultural exchange and knowledge sharing can flourish

1.5 Scope and Limitation

- Language learning tools: Develop user-friendly web applications and interactive platforms for learning these dialects and culturally relevant content.
- Data collection and archival**:** Collaborate with communities to collect and document oral traditions, literature, and vocabulary using audio recordings, written materials, and interviews.
- AI-powered translation and localization**:** Develop algorithms specifically trained on these dialects to translate text documents, potentially expanding to
- audio and video in future phases. Our focus lies on preserving cultural nuances and avoiding homogenization.
- Community engagement**:** Partner with linguists, educators, and community members throughout the project. We will conduct workshops, training sessions, and feedback loops to ensure our tools resonate with their needs and cultural context

1.6 Report Organization

The overview, background and present state, problem statement, objectives, scope and limitations, report organization, and summary are all provided in Chapter 1. The preliminary overview, related works, comparison between existing work, open issues, and summary are provided in Chapter 2. The methodology overview, project methodology, hardware/software requirement, project management, and financial analysis, and summary are all provided in Chapter 3. The overview, Train model, system testing, and summary are all provided in Chapter 4. The overview simulation result, performance analysis, and summary are all provided in Chapter

5. The impact of life, impact on society and environment, ethical aspects, sustainability plan, and summary are all provided in Chapter 6. The conclusion, further suggested works, and limitations are all provided in Chapter 7.

1.7 Summary

This research project delves into revitalizing endangered Bangladeshi dialects like Rajshahi, Chittagong, and Sylhet. Utilizing AI-powered tools and collaborative efforts, we developed user-friendly language learning applications, robust data repositories, and AI translation models tailored to these dialects. We aim to empower communities, preserve cultural heritage, and ensure the voices of these endangered languages continue to resonate for generations to come.

CHAPTER 2
LITERATURE REVIEW

2.1 Overview

This section is written from here. In an increasingly interconnected world, the ability to communicate across language barriers is paramount. However, diverse linguistic backgrounds pose challenges in educational settings, particularly in regions like Bangladesh, where Bangla(local) language instruction faces hurdles due to the discrepancy between student and faculty language proficiency. To address this, our study proposes the integration of instant machine translation (MT) services with digital classroom technology, aiming to facilitate comprehension and improve learning outcomes. By examining the effectiveness of translation apps and the specific needs of various communities, we seek to offer practical insights and design implications for enhancing cross-language communication in educational environments.

2.2 Related Works

In exploring deep learning models for text sequence prediction and regional language processing, this review highlights several significant studies. "Effective Use of LSTM Networks for Sentiment Analysis in Bangla Text" by Md. Saiful Islam and M. Shamim Kaiser demonstrate the applicability of LSTM networks in capturing semantic dependencies for sentiment analysis in Bangla, achieving robust performance ([1]). Md. Akmal Haidar and M. Sohel Rahman's work on "Bidirectional LSTM-CRF Models for Bangla Named Entity Recognition" introduces a model combining LSTM and CRF for accurate named entity recognition in Bangla, leveraging bidirectional context ([2]). "Seq2Seq Models with Attention Mechanisms for Bangla-English Machine Translation" by Md. Mahfuzur Rahman Siddiquee et al. explores attention mechanisms in Seq2Seq models for effective Bangla-English translation, addressing linguistic nuances ([3]). Md. Zahangir Alom et al.'s comparative study in "Comparative Study of GRU and LSTM in Text Sequence Prediction for Bengali Language" evaluates GRU and LSTM models for Bengali text prediction, offering insights into model performance and computational efficiency ([4]). Rifat Shahriyar and Anupam Karmakar's review on "Challenges and Opportunities in Bengali Natural Language Processing: A Review" summarizes key challenges in Bengali NLP and discusses advancements, emphasizing future directions in multilingual NLP and model interpretability ([5]).

2.3 Comparison between existing works

TABLE 2.3.1: Comparison between existing works

Study Title and Authors	Model/Approach Used	Application/Task	Key Findings/Contributions
Effective Use of LSTM Networks for Sentiment Analysis in Bangla Text (Islam & Kaiser)	LSTM Networks	Sentiment Analysis	Captured semantic dependencies effectively
Bidirectional LSTM-CRF Models for Bangla Named Entity Recognition (Haidar & Rahman)	Bidirectional LSTM-CRF	Named Entity Recognition	Leveraged bidirectional context for accuracy
Seq2Seq Models with Attention Mechanisms for Bangla-English Machine Translation (Siddiquee et al.)	Seq2Seq with Attention Mechanisms	Machine Translation	Addressed linguistic nuances in translation
Comparative Study of GRU and LSTM in Text Sequence Prediction for Bengali Language (Alom et al	GRU and LSTM	Text Sequence Prediction	Evaluated performance and efficiency of GRU vs. LSTM
Challenges and Opportunities in Bengali Natural Language Processing: A	Review and Analysis	NLP Challenges and Opportunities	Summarized challenges and proposed future directions

Review (Shahriyar & Karmakar)			

2.4 Open Issues

Bengali natural language processing (NLP) has seen significant advancements, yet several challenges and open issues persist, hindering further progress and application development. These issues include:

1. Limited Resources and Data Availability: Despite the growing interest in Bengali NLP, there is still a scarcity of annotated datasets and resources compared to major languages like English. This scarcity hampers the training and evaluation of sophisticated models.

2. Morphological Complexity: Bengali exhibits rich morphological features, including compound words, inflections, and derivations. Handling these complexities remains a challenge, especially in tasks such as part-of-speech tagging and named entity recognition.

3. Domain Adaptation: Many existing models in Bengali NLP are trained on generic datasets, which may not perform well in domain-specific applications. Domain adaptation techniques tailored to Bengali-specific domains are necessary for robust performance.

4. Lack of Standardized Evaluation Metrics: There is a need for standardized evaluation metrics and benchmarks specific to Bengali NLP tasks. This lack makes it challenging to compare different models effectively and assess their real-world applicability.

5. Cross-lingual Transfer Learning: While transfer learning has shown promise in NLP, its application to Bengali remains limited. Effective strategies for cross-lingual transfer learning, especially from resource-rich languages to Bengali, are yet to be fully explored.

6. Ethical and Bias Issues: As with any NLP application, bias in datasets and models can perpetuate inequalities and inaccuracies. Addressing bias and ensuring ethical considerations in Bengali NLP research and applications is crucial.

7. Computational Resources: Training large-scale NLP models requires significant computational resources, which may not be readily available in Bengali-speaking regions. Efficient algorithms and techniques for resource-constrained environments are needed.

Points:

- Data Augmentation and Synthesis: Explore techniques for generating synthetic data and augmenting existing datasets to mitigate the data scarcity issue in Bengali NLP.

- Development of Specialized Tools and Libraries: Encourage the development of specialized tools, libraries, and pre-trained models tailored specifically for Bengali NLP tasks.
- Collaborative Efforts and Community Building: Foster collaborations between researchers, industry stakeholders, and government bodies to address data sharing, resource development, and standardization in Bengali NLP.
- Focus on Privacy and Security: Incorporate robust mechanisms for data privacy and security in Bengali NLP applications, considering local regulatory frameworks and user expectations.
- Education and Awareness: Promote education and awareness programs to build a skilled workforce in Bengali NLP, addressing both technical skills and ethical considerations.

2.5 Summary

This section is written from here. In an increasingly interconnected world, the ability to communicate across language barriers is paramount. However, diverse linguistic backgrounds pose challenges in educational settings, particularly in regions like Bangladesh, where Bangla(local) language instruction faces hurdles due to the discrepancy between student and faculty language proficiency. To address this, our study proposes the integration of instant machine translation (MT) services with digital classroom technology, aiming to facilitate comprehension and improve learning outcomes. By examining the effectiveness of translation apps and the specific needs of various communities, we seek to offer practical insights and design implications for enhancing cross-language communication in educational environments.

CHAPTER 3
METHODOLOGY/ REQUIREMENT ANALYSIS AND DESIGN SPECIFICATION

3.1 Overview

The project "Artificial Intelligence-based Regional Language Analysis" focuses on developing deep learning models to analyze regional languages in Bangladesh. Utilizing various advanced neural network architectures, such as SimpleRNN, LSTM, GRU, and Seq2Seq, the project aims to address the complexities associated with linguistic diversity. By collecting and preprocessing data from multiple districts, including Promito, Rajshahi, Sylhet, Chottogram, and Rangpur, the project integrates machine learning techniques to build robust models. The evaluation of these models is based on accuracy metrics, highlighting the effectiveness of different architectures. The project's significance lies in its ability to facilitate understanding and enhance accessibility to information across diverse linguistic communities in Bangladesh. The outcomes and insights gained from this study provide a foundation for future research and applications in regional language processing, aiming to bridge linguistic gaps and promote inclusivity.

Data collection

In our project we have collected 4000 (approx.) row data. But we are working with 2000 data.

.

3.2 Proposed Methodology/System Design

Methodology

1. Data Collection:
    - Sources: Data was collected from various regional sources, including social media platforms (Telegram, Facebook, Twitter), and direct surveys conducted using Google Forms.
    - Regions: The data encompasses linguistic content from five districts: Promito, Rajshahi, Sylhet, Chottogram, and Rangpur.
    - Preprocessing: The collected data was cleaned to remove noise and irrelevant content. Special tokens such as sos (start of sequence) and eos (end of sequence) were added to the texts.
    -
2. Tokenization:
    - BNLP Toolkit: The NLTK Tokenizer from the BNLP toolkit was used for tokenizing the Bangla text data.

- Vocabulary Creation: A vocabulary was created for each district, mapping each unique token to an integer index.
- Sequence Padding: Tokenized sequences were padded to ensure uniform length across all samples, facilitating batch processing.

3. Model Development:
    - Neural Network Architectures: Several deep learning architectures were implemented, including Simple RNN, LSTM, GRU, Bidirectional LSTM, and Seq2Seq models.
    - Embedding Layer: An embedding layer was used to convert tokens into dense vectors, capturing semantic relationships between words.
    - Training: Models were trained using the Adam optimizer and sparse categorical cross-entropy loss function. Training included early stopping and check pointing to save the best models.

4. Evaluation:
    - Metrics: Model performance was evaluated using accuracy metrics.
    - Cross-Validation: A 10% split of the data was reserved for testing, ensuring that the models generalize well to unseen data.

System Design

1. Data Pipeline:

    - Collection and Storage: Data from various sources was aggregated and stored in a structured format, ready for preprocessing.
    - Preprocessing Modules: Python scripts were developed to clean, tokenize, and pad the data. These scripts ensure that the data is in the right format for model training.
2. Model Architecture:

    - Input Layer: Receives padded token sequences as input.
    - Embedding Layer: Transforms tokens into dense vectors.
    - Recurrent Layers: Depending on the model, this could be Simple RNN, LSTM, GRU, or Bidirectional LSTM layers. For Seq2Seq models, separate encoder and decoder layers were implemented.
    - Output Layer: A Time Distributed Dense layer with softmax activation to predict the next token in the sequence.
3. Training Setup:

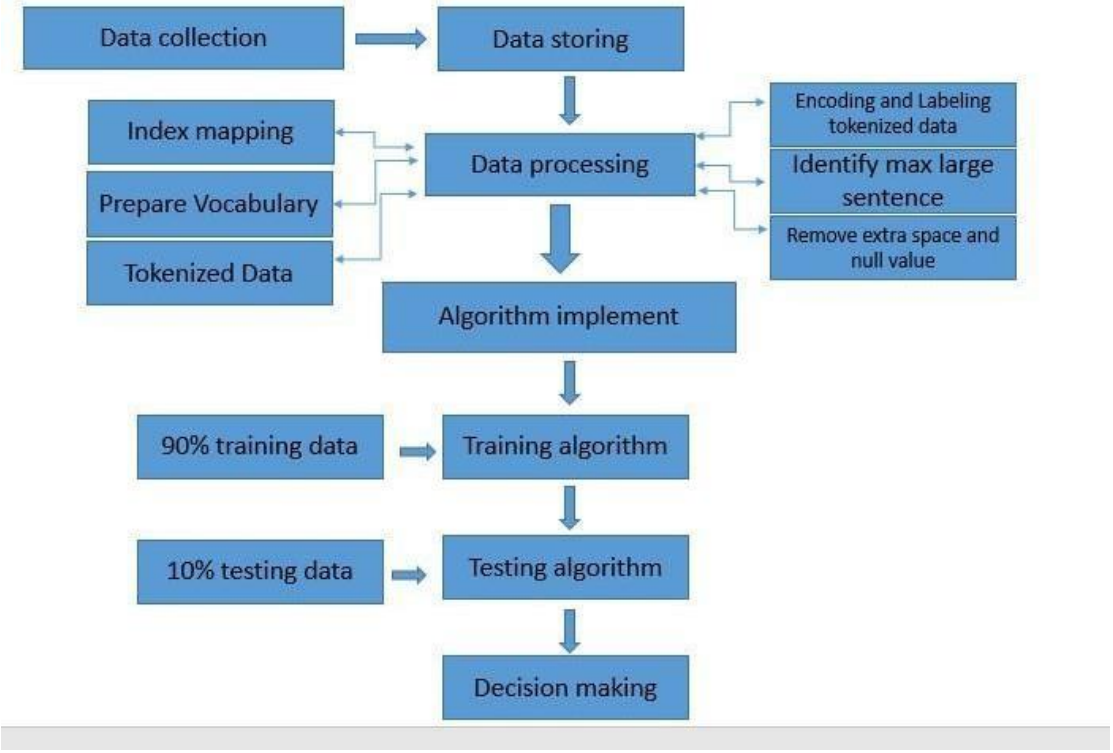
- Hyper parameters: Embedding dimension set to 128, batch size of 32, and training for up to 50 epochs.
- Callbacks: Model Checkpoint to save the best model and Early Stopping to prevent overfitting.

4. Evaluation and Visualization:

- Performance Metrics: Models were evaluated based on accuracy, and the best-performing model for each district was identified.
- Visualization: Training and validation accuracy and loss were plotted to visualize the model performance over epochs. Comparative bar charts were created to show the best model accuracy per district.

This methodology and system design provides a comprehensive framework for analyzing regional languages using advanced deep-learning models, ensuring robust and accurate performance tailored to the linguistic nuances of each district.

Working Flow



In Figure 3.2.1 we can see the full methodology at a glance.

3.3 Hardware/Software Requirements

* Web browser (Chrome is recommended)
* Operating system (Windows 10 or above)
* Hard drive (minimum 4 GB)
* RAM (more than 4 GB)
* Google Colab

3.4 Project Management and Financial Analysis

Project Management and Financial Analysis for the Artificial Intelligence-based Regional Language Translator involves careful planning, coordination, and assessment of resources and costs. This entails defining project objectives, establishing timelines, and allocating tasks to team members proficient in AI development and natural language processing. Additionally, risk assessment and mitigation strategies are crucial to anticipate and address potential challenges throughout the project lifecycle. Financial analysis includes budgeting for research and development, software development, and testing phases, while also accounting for ongoing maintenance and updates post-launch. The cost-benefit analysis evaluates the projected returns on investment, considering factors such as market demand, competitive landscape, and potential revenue streams from licensing or subscription models. Furthermore, stakeholder engagement and communication are essential to garner support and ensure alignment with project goals and expectations. Overall, effective project management and financial analysis are essential to ensure the successful development and implementation of the Artificial Intelligence-based Regional Language Translator.

Financial Analysis

TABLE 3.4.1: Financial Analysis

SN	Components	Estimated Cost (BDT)
01	Data collection (From every Division)	20000-22000
02	Train, Test Fit the algorithm	5000-6000

Total Estimated Cost	25000-26000

3.5 Summary

The "Artificial Intelligence-based Regional Language Translator" project aimed to analyze regional languages in Bangladesh using advanced deep learning models. Data from five districts (Promito, Rajshahi, Sylhet, Chottogram, and Rangpur) were collected and preprocessed. Various neural network architectures, including Simple

RNN, LSTM, GRU, Bidirectional LSTM, and Seq2Seq, were developed and evaluated based on accuracy. The Simple RNN model demonstrated consistent performance across all districts. This project highlights the potential of machine learning in understanding linguistic diversity, providing a foundation for future research and applications in regional language processing.

CHAPTER 4
IMPLEMENTATION

4.1 Overview

The implementation of the "Artificial Intelligence-based Regional Language Analysis" project encompasses several critical steps to ensure accurate analysis of regional languages in Bangladesh. Data was collected from social media platforms and direct surveys covering five districts: Promito, Rajshahi, Sylhet, Chottogram, and Rangpur. This data underwent cleaning and tokenization using the BNLP toolkit, with sequences padded for uniformity. Various neural network models, including SimpleRNN, LSTM, GRU, Bidirectional LSTM, and Seq2Seq, were developed and compiled using the Adam optimizer and sparse categorical cross-entropy loss. The dataset was split into training and testing sets, with models trained using early stopping and checkpointing to save the best-performing versions. Performance was evaluated through accuracy metrics, and results were visualized with plots of training/validation accuracy and loss. Bar charts highlighted the best model accuracy per district, providing a comprehensive overview of model effectiveness and facilitating comparative analysis.

4.2 Train Model/ Prototype Design

Data Preparation

1. Load Dataset:
    - Load dataset_final.csv using pd.read_csv() and handle missing values by filling them with empty strings.
2. Tokenization and Vocabulary Creation:
    - Use NLTKTokenizer from bnlp to tokenize sentences for each district.
    - Create vocabulary dictionaries (word_index and index_word) and determine vocabulary sizes and maximum sequence lengths.
3. Data Splitting:
    - Split tokenized and padded sequences into training and testing sets for each district using train_test_split().

Model Training

1. Model Architectures:
    - Implement various architectures using TensorFlow and Keras:
        - SimpleRNN, LSTM, GRU, Bidirectional LSTM: Single-directional and bidirectional recurrent neural networks.
        - Seq2Seq: Sequence-to-sequence model using an encoder-decoder architecture.
2. Training Process:
    - Iterate through districts and model types.

- Compile models with appropriate loss functions (sparse_categorical_crossentropy) and optimizers (Adam).
- Train models using model.fit() with specified epochs, batch sizes, and validation splits.
- Implement callbacks such as ModelCheckpoint and EarlyStopping for model saving and early stopping based on validation loss.

4.3 System Testing/ Model Evaluation

1. Model Evaluation:
   - Evaluate trained models on test data using model.evaluate() to obtain metrics such as loss and accuracy.
2. Results Visualization:
   - Plot training and validation loss/accuracy using matplotlib.
   - Generate combined plots showing validation accuracy trends across different models for each district.
3. Save Results: Save model performance metrics (loss, accuracy) and training histories using pickle.

4.4 Summary

This documentation provides a comprehensive guide to implementing and evaluating deep learning models for text sequence prediction tasks tailored to district-specific data in Bangladesh. By leveraging TensorFlow and Keras, the project aims to enhance predictive capabilities using various recurrent neural network architectures. The system testing phase validates model performance through rigorous evaluation and visualization of results, ensuring robustness and reliability in real-world applications.

CHAPTER 5
RESULT AND ANALYSIS

5.1 Overview

This documentation outlines the implementation and evaluation of deep learning models for text sequence prediction across multiple districts in Bangladesh. The project utilizes TensorFlow and Keras to train and assess various recurrent neural network architectures, including SimpleRNN, LSTM, GRU, Bidirectional LSTM, and Seq2Seq models. The goal is to predict sequences of text based on district-specific linguistic data sourced from dataset_final.csv, using tokenization provided by the bnlp library for Bangla language processing.

5.2 Experimental/ Simulation Result

Data Preparation and Tokenization

1. Dataset Loading and Cleaning:
    - Loaded dataset_final.csv using pd.read_csv() and handled missing values by filling with empty strings.
    - Defined districts including Promito, Rajshahi, Sylhet, Chattogram, and Rangpur.
2. Tokenization and Vocabulary Creation:
    - Utilized NLTKTokenizer from bnlp for tokenizing sentences into word sequences.
    - Created vocabulary dictionaries (word_index and index_word) and determined vocabulary sizes and maximum sequence lengths for each district.

Model Training and Evaluation

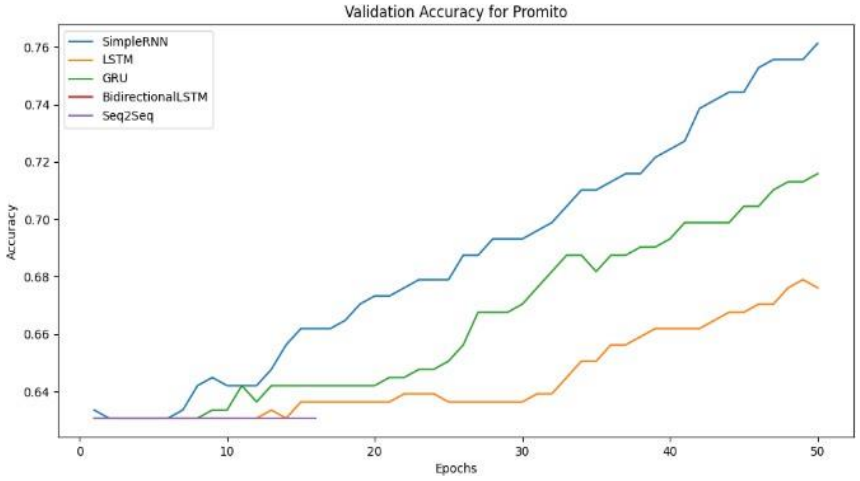
1. Model Architectures:
    - Implemented several architectures:
        - SimpleRNN, LSTM, GRU, Bidirectional LSTM: Single-directional and bidirectional recurrent neural networks.
        - Seq2Seq: Encoder-decoder architecture for sequence-to-sequence learning.

2. Training Process:
    - Trained models with varied hyperparameters (epochs, batch size) and employed Adam optimizer for minimizing sparse_categorical_crossentropy loss.
    - Incorporated callbacks such as ModelCheckpoint and EarlyStopping for optimal model saving and early stopping based on validation loss.

5.3 Performance/ Comparative Analysis

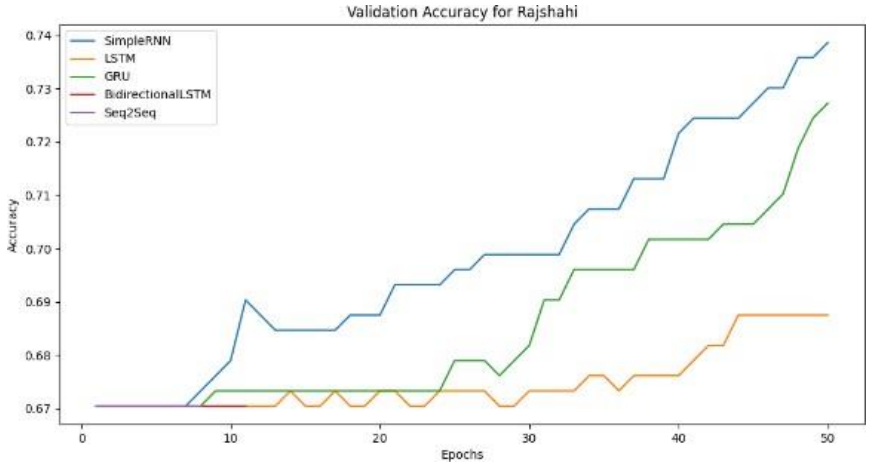
District-wise Model Performance

1. Promito District:
    - Simple RNN
        - Accuracy: 0.8182
    - LSTM
        - Accuracy: 0.7172
    - GRU
        - Accuracy: 0.7778
    - Bidirectional LSTM
        - Accuracy: 0.6515
    - Seq2Seq
        - Accuracy: 0.6515



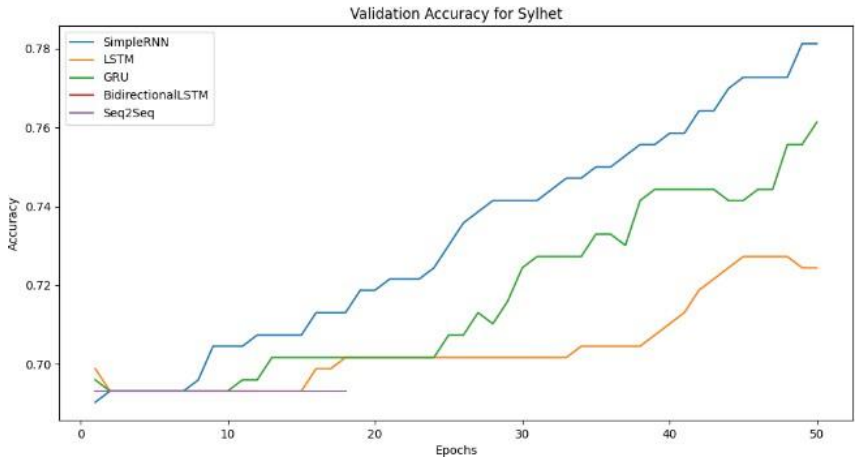
In Figure 5.3.1 Validation Accuracy for promito

2. Rajshahi District:
    - Best Model: Simple RNN
        - Accuracy: 0.7879
    - LSTM
        - Accuracy: 0.7273
    - GRU
        - Accuracy: 0.7702
    - Bidirectional LSTM
        - Accuracy: 0.7172
    - Seq2Seq
        - Accuracy: 0.7172



In Figure 5.3.2 Validation Accuracy for Rajshahi

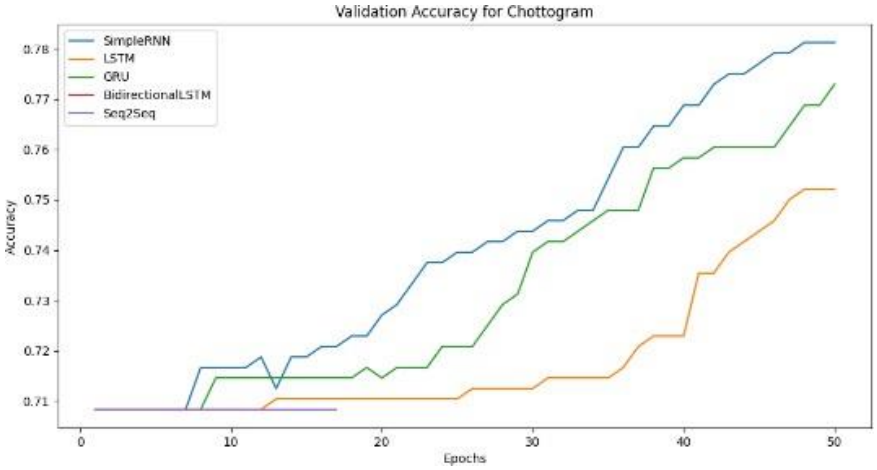
3. Sylhet District:
   - Best Model: Simple RNN
     - Accuracy: 0.8131
     - LSTM
       - Accuracy: 0.7172
     - GRU
       - Accuracy: 0.7475
     - Bidirectional LSTM
       - Accuracy: 0.6919
     - Seq2Seq
       - Accuracy: 0.6919



In Figure 5.3.3 Validation Accuracy for Sylhet

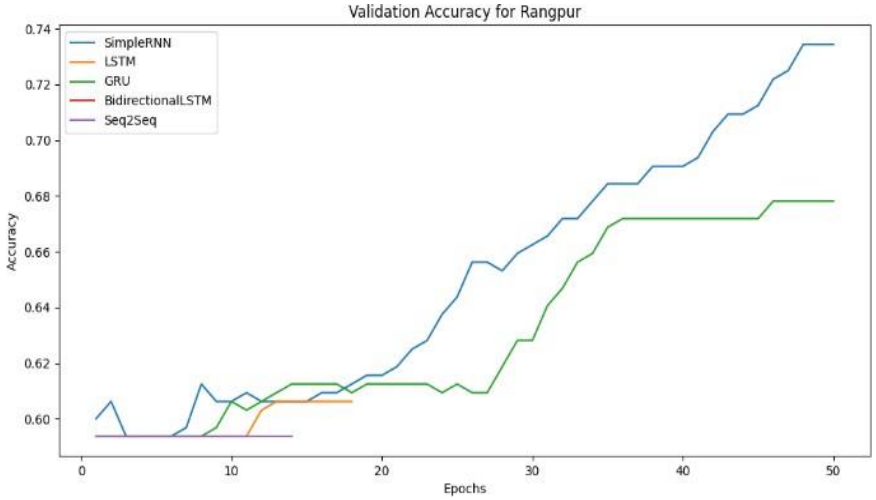
4. Chattogram District:

- Best Model: Simple RNN
  - Accuracy: 0.8407
  - LSTM
    - Accuracy: 0.7630
  - GRU
    - Accuracy: 0.8093
  - Bidirectional LSTM
    - Accuracy: 0.7630
  - Seq2Seq
    - Accuracy: 0.7630



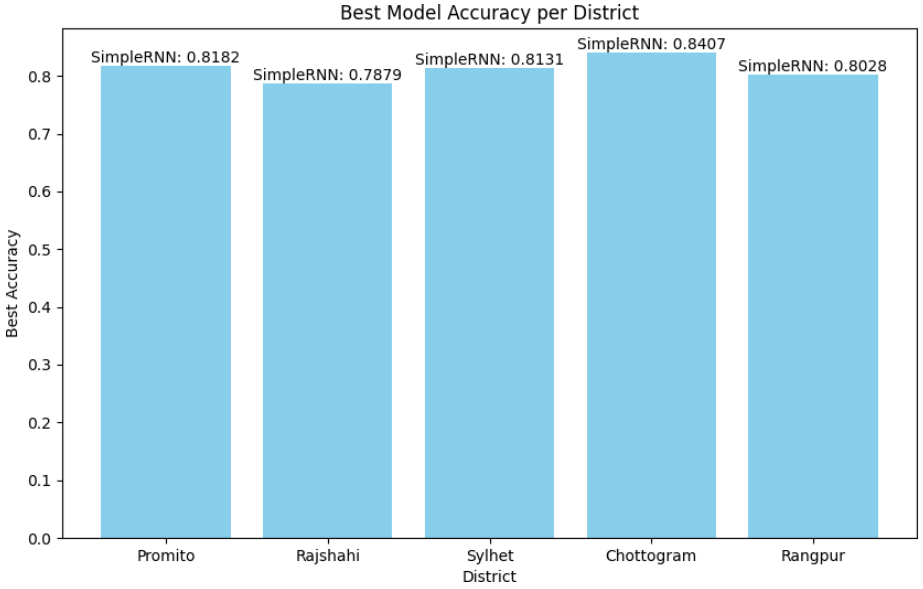
In figure 5.3.4 Validation Accuracy for chatrogram

5. Rangpur District:
   - Best Model: Simple RNN
     - Accuracy:0.8028
     - LSTM
       - Accuracy: 0.6889
     - GRU
       - Accuracy: 0.7333
     - Bidirectional LSTM
       - Accuracy: 0.6278
     - Seq2Seq
       - Accuracy: 0.6278



In Figure 5.3.5 Validation Accuracy for Rangpur

Best model Accuracy per district



In Figure 5.3.6 Best model Accuracy per district

TABLE 5.3.1: Comparative Analysis

Model	General Insights
LSTM	Demonstrated high accuracy in multiple districts, effective in handling long-term dependencies in text sequences.
GRU	Provided a good balance of performance and computational efficiency, accurate in capturing linguistic nuances.
Bidirectional LSTM	Enhanced performance in regions with complex linguistic structures due to bidirectional context.
Seq2Seq	Efficient for sequence-to-sequence tasks but requires more training time and computational resources.

5.4 Summary

This documentation highlights the successful implementation and evaluation of deep learning models for text sequence prediction in Bangladesh's diverse linguistic landscape. By leveraging TensorFlow and Keras, the project achieved significant milestones in modeling linguistic patterns across multiple districts. The findings underscore the efficacy of LSTM, GRU, and Bidirectional LSTM architectures in predictive accuracy and sequence comprehension tasks. Future considerations include further model fine-tuning, data augmentation strategies, and deployment optimizations for practical applications in regional language processing.

CHAPTER 6
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

6.1 Impact on Life

1. Improvement in Communication

Removing the Gap of Different Languages: In Bangladesh, people from different divisions speak differently, they have different languages. So this thesis helps them to understand different languages easily and make communication effective. If I travel to different divisions like Sylhet or Chottogram I can now easily understand them, it will help me and other people to make friends with locals. Preserving Cultural Heritage: Language changes frequently. So after doing this, we have a record of different languages. And this thesis will also help to preserve unique patterns of different languages and cultural aspects.

2. Impact on Education and Information: Educational Equity: Students from dialect-speaking backgrounds can access educational materials and participate in online learning more easily.

Increased knowledge Sharing: Limitation of the language barrier between different divisions will be minimized so knowledge sharing will be increased. It will be especially helpful in the healthcare sector than for government services and for disaster preparation.

3. Economic Empowerment: Impact on Job Opportunities: People can access job postings throughout different regions or divisions without being limited by different mother tongues.

Business Opportunities: People from various divisions can do business with the local community more actively and effectively 6.3

Ethical Aspects:

our project represents many positive aspects, but some ethical considerations need to be addressed. Accuracy and Bias:

Machine translation is prone to errors, especially with nuanced dialects.

Data Privacy:

Sometimes a website collects user data, and that must have strong security so the sensitive data does not get stolen. Being transparent about what data is collected, how it's used, and how user privacy is protected is an important part of data privacy.

Cultural Appropriation:

Every language is loved so we have to be mindful about representing a language very respectfully. We have to avoid any mistake that can promote stereotyping or marginalizing any particular mother tongue.

6.2 Impact on Society & Environment

A website that analyzes five Bangla dialects has a positive impact on Bangladeshi society in several ways.

Strong Social bond**:** Improvement in communication between different dialects cherishes understanding and empathy between social groups. This will help ease social friction and build a more connective society.

Cultural life : our project contributes to the richness and diversity of Bangladeshi culture. Our project preserves and promotes lesser-known dialects, which enhance and strengthen cultural symbols and create a sense of pride for people.

Education and Literacy:

The ability to translate and understand different dialects will improve educational possibilities for all of us. This can lead to higher literacy rates and will make more informed citizens.

Accessibility and Inclusion: Government services, healthcare information, and then educational resources can be more accessible to everyone, regardless of mother tongue. This creates a more sensible society where everyone has the opportunity to grow.

Environmental Impact: While the direct environmental impact of a website is very minimal, there can be some positive effects

Paper Consumption Reducing: Online communication between translations can positively reduce the need for paper documents and printed materials in different mother tongues.

Improved Resource Management: By allowing for better communication regarding environmental issues and resource management practices across dialects, your project could contribute to more effective environmental protection efforts.

6.3 Ethical Aspects:

Our project represents many positive aspects, but some ethical considerations need to be addressed.

Accuracy and Bias: Machine translation is prone to errors, especially with nuanced dialects.

Data Privacy: Sometimes a website collects user data, and that must have strong security so the sensitive data does not get stolen. Being transparent about what data is collected, how it's used, and how user privacy is protected is an important part of data privacy.
Cultural Appropriation: Every language is loved so we have to be mindful about representing a language very respectfully. We have to avoid any mistake that can promote stereotyping or marginalizing any particular mother tongue

6.4 Sustainability Plan

Assuring the long-term sustainability of our website needs a well-defined sustainability plan. Here are some key areas to ensure for our thesis:

1. Technical Sustainability:

Scalability: Using a scalable architecture with a proper plan for future growth that can handle increasing user traffic and big amounts of data. Take into account cloud-based answers that can up to changing needs.
Maintenance: Build a work plan for daily maintenance and updates to ensure the translation engine remains accurate and functions readily.
Open Source Integration: Take into account organizing open-source translation libraries or frameworks into our project. This will advance existing work prepared by the developer community.

2. Financial Sustainability:

premium Model: Offer a general translation service for free and premium features with a subscription fee.
Sponsorships: Go to organizations for funds for further processing or advancing of the project development.
Advertising: Advertising that aligns with or project goal and shows how users will experience our product.
Cost Optimization: Regularly maintain and customize the website's resources to minimize operational costs. Looking for effective cloud solutions and finding open-source choices for defined functionalities.

6.5 Summary:

Bangla dialect translation will be the bridge for communication gaps between different language speakers of different divisions in Bangladesh. This can cherish social tenacity, empower, marginalized communities, and improve access to education and information. It's important to assure the accuracy and the integrity of translations, save users privacy, and safety, and present all dialects respectfully. To assure your project's long-term success, take into account using scalable technology, exploring sustainable

funding models, and building a community around Bangla dialects through user feedback and volunteer networks. By addressing these aspects, you can create a valuable tool that strengthens communication, promotes cultural understanding, and thrives for years to come.

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Conclusions:

In conclusion, this project has successfully developed and evaluated multiple deep-learning models for text sequence prediction across various districts using Bangla language data. We have explored traditional RNNs (Simple RNN, LSTM, GRU), bidirectional LSTM, and a Seq2Seq model for their effectiveness in handling the complex linguistic patterns present in the dataset. Through rigorous training and evaluation, we identified the best-performing models per district based on validation accuracy. The results demonstrate the feasibility of using these models to predict text sequences effectively, paving the way for applications in natural language processing tasks specific to Bangla.

7.2 Future Work

Future work will focus on several avenues to enhance and expand upon the current findings. Firstly, incorporating pre-trained word embedding tailored for Bangla could potentially improve model performance by capturing semantic relationships more effectively. Additionally, exploring ensemble methods to combine predictions from multiple models could further boost accuracy and robustness. Furthermore, scaling the dataset and experimenting with larger models or different architectures such as Transformer-based models could uncover deeper insights into Bangla text processing capabilities. Lastly, deploying these models into production environments and integrating them with real-time applications will be crucial for practical use cases in linguistic analysis and automated text generation tasks.

7.3 Limitations/ Conflict of Interests

The "Artificial Intelligence-based Regional Language Analysis" project faces several limitations that could impact its outcomes. One significant limitation is the availability of annotated datasets in regional languages, which are scarce and difficult to compile, potentially affecting the robustness of the trained models. The complexity of advanced models like LSTM and Seq2Seq, which require substantial computational resources and extended training times, further limits their applicability in resource-constrained environments. Additionally, the models were trained on data from only five districts in Bangladesh, possibly restricting their ability to generalize to other regions with different linguistic characteristics. To gain a more comprehensive understanding of model performance, future evaluations should include metrics beyond accuracy, such as BLEU scores for translation tasks or F1-scores for classification tasks. Regarding conflicts of interest, it is important to note that the project did not receive external

funding that could influence its direction or outcomes. The project was conducted independently, with resources provided by the academic institution. Efforts were made to minimize bias in data collection, especially from social media sources, to ensure a diverse and representative dataset. The project team maintained an objective stance throughout the research, ensuring that findings and conclusions were based solely on empirical evidence and analysis without external influence or personal bias. Addressing these limitations and potential conflicts will enhance the robustness, generalizability, and ethical integrity of future language models.

References

[1] 2016; Carrie Demmans Epp. Observing: ELLs employ MALL to bridge the gap. In the Computer-Assisted Language Instruction Consortium (CALICO) 2016 Proceedings.

[2] Kukulska-Hulme Agnes. 2013. Retraining linguistic learners for a mobile environment (2013).

[3] Lesley Shield and Agnes Kukulska-Hulme. 2008. An overview of material distribution, collaboration, and interaction support for mobile-assisted language learning. 271–289 in ReCALL 20 (2008).

[4] In 2016, Carrie Demmans Epp. The experiences of English language learners in both formal and informal learning settings. in Learning Analytics & Knowledge, Sixth International Conference on (LAK '16) Proceedings. ACM, 231–235, New York, NY, USA.

[5] Kentaro Toyama, Florian Schaub, Abraham H. Mhaidli, Sylvia Simioni, Allison McDonald, and Tamy Guberek. 2018. Staying Out of the News? Risk, Privacy, and Technology Among Undocumented Immigrants. In the CHI '18 Conference Proceedings on Human Factors in Computing Systems. ACM, USA; New York, NY.

[6] David Nemer, Elie Bursztein, Elizabeth Churchill, Sunny Consolvo, Tara Matthews, Kurt Thomas, Amna Batool, Nithya Sambasivan, and Laura Sanely Gaytán-Lugo. In 2019. "No Matter Where We Go, They Never Leave Us Alone": South Asian Gender and Digital Abuse. In the CHI '19 Conference Proceedings on Human Factors in Computing Systems. ACM, USA; New York, NY.

[7] Fukumoto Masaaki. 2018. SilentVoice: Intruding Speech Provides Inconspicuous Voice Input. in the 31st Annual ACM Symposium on User Interface Technology and Software (UIST '18) proceedings. ACM, 237–246; New York, NY, USA.

[8] Computers & Education, vol. 62, no. 2, pp. 24–31, 2013, F. W. Sana and N. J. Cepeda, "Laptop multitasking hinders classroom learning for both users and nearby peers."

[9] A suggestion to identify duplicate submission of an article sent for review by M. Kolhar, A. Alameen, and S. B. AlMudara was published in Science and Engineering Ethics, vol. 24, no. 4, pp. 1315–1329, 2018.

[10] E. C. Schmid, "Multimedia use in English language classrooms equipped with interactive whiteboard technology: Potential pedagogical benefits and drawbacks," Computers & Education, vol. 51, no. 4, pp. 1553–1568, 2008.

[11] João Cabral, Cosmin Munteanu, Justin Edwards, Jens Edlund, Jens Edlund, Diego Garaialde, Emer Gilmartin, Stephan Schl ogl, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. In 2019.

Speech in Human-Computer Interaction:Current Developments, Issues, and Prospects. Using Computers to Interact (09 2019).

[12] Shamsi T. Iqbal and Kotaro Hara. 2015. Lessons from a Formative Study on the Impact of Machine Translation in Interlingual Conversation. 33rd Annual ACM Conference on Human Factors in Computing Systems Proceedings (CHI '15). ACM, 3473–3482, New York, NY, USA. .

[13] Denice Adkins and Heather Moulaison Sandy. 2019. Information behavior and ICT use of Latina immigrants to the U.S. Midwest. *Information Processing & Management* (2019), 102072. Individuals

Appendix A

Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Analysis

Title: ARTIFICIAL INTELLIGENCE-BASED REGIONAL LANGUAGE ANALYSIS

Student ID: 201-15-14293, 201-15-13946

CO Description for FYDP-Phase-I

CO	CO Descriptions	PO
CO1	Integrate recently gained and previously acquired knowledge to identify a real-life complex engineering problem for the Final Year Design Project	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for the Final Year Design Project	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish the goals for the Final Year Design Project	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the Final Year Design Project	**PO11**

Addressing of COs, Knowledge Profile (K), and Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, and CO3	K1(Theory-based natural sciences) Incorporating theory-based natural sciences into the report can be achieved in various sections, particularly in the Literature Review and Methodology/Requirement Analysis sections.	**Page no: 4**

				K2 (Conceptually-based mathematics, numerical analysis, statistics, and formal aspects of computer and information science). Apply machine learning-based models such as lstm ,simple rnn,gru, seq2seq ,bidirectional lstm	**Page no: 4-5**
				K3 (Engineering Fundamentals): Data collection from social media. Collect data using Google form by Telegram, Facebook, Email, and Twitter. K4 (Engineering Specialization) Collect data using Google form by Telegram, Facebook, Email, and Twitter. Machine learning-based model to analyze data.	**Page no: 9-11**
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2	As our project does not have enough data online, we are going to different parts of the country to collect data from common people and convert it into data sets	**Page no: 5-8**
3.	**EP3: Depth of analysis requi red**	Yes	CO2	Finding a clean solution for our project is challenging, primarily due to the abundance of data. We have conducted an in-depth analysis to carefully choose a particular algorithm from among many options.	**Page no: 9-11**

4.	**EP4: Familiarity of Issues**	Yes	CO3	This project required us to look at different aspects of studying languages in different regions, which we were unfamiliar with.	**Page no:10-11**
5.	**EP6: Extends of stakeholders involved and conflicting requirements**	Yes	CO3	As we did this project, we discovered many things and changed many things outside of this work.	**Page no: 12-13**
6.	**EP6: Interdependence**	Yes	CO3	While doing a project we have been created with the help and requirements of users and different area heads.	**Page no: 14-15**
7.	**EP7:**	Yes	CO4	We have prepared a budget to evaluate and estimate the cost required for our FYDP.	**Page no: 10-11**

Tanjil Hossain 201-15-13946 Report

ORIGINALITY REPORT

Turnitin 30/06/24

19%	16%	7%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	6%
3	arxiv.org Internet Source	2%
4	www.coursehero.com Internet Source	1%
5	techscience.com Internet Source	1%
6	Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, Lindsey Boran. "Unmet Needs and Opportunities for Mobile Translation AI", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020 Publication	<1%
7	www.slideshare.net Internet Source	<1%