

MODELLING OF BANGLA REAL-WORD ERROR CORRECTION

BY

RUDRA SARKER UTSHA

ID: 173-15-10422

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

Supervised By

MD. TAREK HABIB

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

PROFESSOR DR. MD. ISMAIL JABIULLAH

Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This research project titled “**Modelling of Bangla Real Word Error Correction**”, submitted by **Rudra Sarker Utsha** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the 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 January 2022.

BOARD OF EXAMINERS



Professor Dr. Touhid Bhuiyan

Chairman

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Dr. Fizar Ahmed

Internal Examiner

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Ms. Nusrat Jahan

Internal Examiner

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

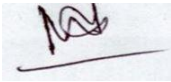
We hereby declare that this project has been done by us under the supervision of **Md. Tarek Habib, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Professor Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Submitted by:



Rudra Sarker Utsha
ID: 173-15-10422
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, I express my sincerest appreciation and gratitude to almighty God for His divine blessing that makes me possible to complete the final year project/internship successfully.

I am obliged and wish my profound gratitude to **Md. Tarek Habib, Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “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 them at all stages have made it possible to complete this project.

I would like to express our heartiest appreciation to **Professor Dr. Md. Ismail Jabiullah**, Professor, Department of CSE, for his kind help to finish my project and also to other faculty members and the staff of the CSE department of Daffodil International University.

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

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

ABSTRACT

This research project “**Modelling of Bangla Real Word Error Correction**” is a language model for finding real-word errors in a Bangla sentence and providing correction on the error word. This topic is now very relevant in the Natural Language Processing sector as it is now a topic of huge interest. The syntactical and grammatical rules in Bangla are rather complex, which poses trouble in handling the language. Words can be obscure where the meaning is dependent on the context. In this project, we proposed a model with Bidirectional LSTM model, which is short for Long Short-Term Memory model. LSTM is a RNN (Recurrent Neural Network) architecture that can not only process single data point but an entire sequence of data. Firstly, the Trigram sequence was created to get context out of a sequence, and fed into the LSTM model. Since the Bidirectional LSTM model remembers the forward as well as the backward relationship of a sequence, it can have a better understanding of the context of a Bangla sentence. After training the model and implementing it to detect and provide correction of a real word error we got an accuracy of 74.450% on the test dataset. But in predicting the next word from the sentence context it was even more successful with 85.47% accuracy. This proposed model was tested in many ways after implementation and it works successfully in both detecting and correcting the real word error.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	iii
Acknowledgements	iv
Abstract	v
List of Figures	ix
List of Tables	ix
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Objectives	2
1.4 Expected Outcome	3
1.5 Report Layout	3
CHAPTER 2: BACKGROUND	5-8
2.1 Preliminaries/Terminologies	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	7

2.4 Scope of the Problem	8
2.5 Challenges	8
CHAPTER 3: Research Methodology	9-15
3.1 Research Subject and Instrumentation	9
3.2 Data Collection Procedure/Dataset Utilized	9
3.3 Statistical Analysis	10
3.4 Proposed Methodology/Applied Mechanism	11
3.5 Implementation Requirements	15
CHAPTER 4: Experimental Results and Discussion	16-17
4.1 Experimental Setup	16
4.2 Experimental Results & Analysis	16
4.3 Discussion	17
CHAPTER 5: Impact on Society, Environment and Sustainability	18-19
5.1 Impact on Society	18
5.2 Impact on Environment	18
5.3 Ethical Aspects	19
5.4 Sustainability Plan	19

CHAPTER 6: Summary, Conclusion, Recommendation, and Implication for Future Research	20-21
6.1 Summary of the Study	20
6.2 Conclusions	20
6.3 Implication for Further Study	21
APPENDIX	25
Appendix A: Project Reflection	25
REFERENCES	22-24

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: Most Frequent Words in Corpus	10
Figure 3.1.2: Text Processing and training dataset	12
Figure 3.2.1: Neural Network Architecture	14
Figure 3.2.2: System Architecture for Bangla Real Word Error Detection	15

LIST OF TABLES

TABLES	PAGE NO
Table 1: Terminologies	5
Table 2: Accuracy	16
Table 3: Input/Output Assessment	17

CHAPTER 1

INTRODUCTION

1.1 Introduction

The written form of any language has been the single most important means of communicating and storing information from the early etchings in clay to the world of digital access that we now enjoy. As our communication with the text increased day by day with the development of technology, a large number of people share their ideas across the internet in Bangla. Detecting and correcting a real-word error in a Bangla sentence is an important and complex task in Natural Language Processing (NLP) that helps predict the correct word to complete the sentence logically. In-text documentation, there are two types of spelling errors, non-word spelling error, and real-word spelling error. Non-word error occurs when the word itself does not exist in the dictionary and has no real meaning. Unlike non-word error, real-word error occurs in a sentence when the word is correct in grammar and is a meaningful word in the dictionary but does not match the sentence.

The Bangla language has intricated orthographic rules as well as many composite Grammatical requirements that are difficult to uphold. Detecting a spelling error in Bangla is difficult on its own, checking real-word error in the Bangla sentence conveys supplementary strain. Whereas writing or discussing in our daily work life, there might be a real-word error while the word is accurate in the dictionary but alters the original meaning of the sentence which should not be unheeded. An example to help us understand such error would be as " ফুটবল একটি জনপ্রিয় খুলি ". Here, even if the word " খুলি " is a correct word it is still incorrect. More appropriately in the sentence mentioned " খুলি " can be changed to " খেলা " and in the sentence, it is suitable and sensible. In an effort to solve this kind of problem a Deep Learning approach can give us a solution. In this project, a tri-gram word sequence and Long Short-Term Memory (LSTM) model were implemented and tested to solve the real-word error-correction problem. In spite of being an important topic of great interest today, no Satisfactory research has been done in the Bangla language on finding and correcting a real word error in sentences. This project proposed a method to address the real-word error detection and correction in Bangla sentences

using some popular NLP techniques and LSTM recurrent neural networks. For this project, Bangla newspaper article was collected and processed to make a trainable corpus. The model architecture was trained in small batches but the processed corpus was not reusable for further work. This proposed approach helps to recompence the inconsistencies of a word in a Bangla sentence with a word with a high probability to match the context of the sentence. The whole paper is arranged as follows. Category 2 includes existing jobs. Section 3 represents the proposed process. Part 4 highlights used tools and technology and section 5 deals with testing the result.

1.2 Motivation

Although a significant number of Deep Learning model has been proven to be very capable of processing and finding structures in natural language, there hasn't been much work done in real-word error detection in the Bangla language. There has been a lot of work done in non-word error detection and correction but most of them were non-contextual and very few of them tackled the problem of real-word error detection and correction. Many valuable research projects relevant to this research have been led in English as well as other languages, but in Bangla, we remain far behind. The result of this research can very well be used in various practical sectors such as Bangla article writing, essay, and blog writing as well as books and other writing sectors were making small mistakes in writing is inevitable. This research is focused on Bangla news article writing and real-word error in Bangla news articles. To help provide an error-free writing experience and shorten the time of revision before publishing an article is the motivation of this research project.

1.3 Objectives

- To check for any real word error in a Bangla sentence.
- To get a context of a sentence.
- To predict next word from the previous context.

- To get an assumption about the sentence context.
- To detect the error word of a sentence.
- To provide correction for the error word.
- To save time and give a smooth writing experience.
- To review a Bangla article for grammatical and contextual word error.

1.4 Expected Outcome

1. A satisfactory result on detection of real-word error.
2. A satisfactory set of word suggestions for an error word.
3. An implementation that is not too complex.
4. Detection of real-word error in Bangla sentence in a news article.
5. People will not need to worry about silly mistakes in their writings.
6. It will save their time, complexities, money, and other things.

1.5 Report Layout

Chapter 1: Introduction

Motivation, objectives and the expected outcome of the project have been discussed in this chapter. The report layout has been mentioned in the last part of this chapter.

Chapter 2: Background

Foundation conditions of my project have been talked about here. Related work, contrasting and other sites, the extent of the issue and difficulties of the project are clarified here.

Chapter 3: Research Methodology

This chapter discusses about the research methodology. Preprocessing of dataset, details of dataset, method used to prepare data for training and testing. And mainly the n-gram sequence and LSTM model and implementation after training.

Chapter 4: Experimental Results and Discussion

This chapter consists of all the testing and experimental results and what the result says in detail.

Chapter 5: Impact on Society, Environment and Sustainability

Impact on Society, Environment, and Sustainability are the contents of this chapter.

Chapter 6: Summary, Conclusion, Recommendation, and Implication for Future Research

It is the last chapter where a summary of the research, conclusion, and future scope of the research have been discussed.

CHAPTER 2

BACKGROUND

2.1 Preliminaries/Terminologies

This is a Deep Learning based Natural Language Processing project, where we tried to find the relations and grammatical structure of Bangla Language using Machine Learning tools. In Table 1 terminologies for this project is listed.

Table 1: Terminologies

Acronym	Definition
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Network

2.2 Related Work

A great deal of important examination works applicable to this exploration have been led in English just as some other languages consistently. However, at that point once more, slight headway is made in the Bangla language. B. Kundu et al. utilized a characteristic language age model for revising Bangla linguistic mistakes [1]. However, it doesn't tackle the issue real-word mistake. For recognition of a semantic mistake in Bangla sentences, K. M. A. Hasan et al. proposed a model [2]. Their method is for easygoing sentences as Subject, Object and Verb dependent on subject-action word and article action word

connection. Z. Yuan and T. Briscoe utilized neural machine interpretation structure for linguistic blunder remedying for English sentences [3]. They applied RNN search model that includes bidirectional RNN and an encoder to gain data. P. Samanta and B. B. Chaudhuri proposed a technique that attempts to distinguish mistake noticing bigram and trigram established by and large left and right neighbor of up-and-comer word and creates idea as per the positions of components of disarray set [4]. Despite the fact that their model showed great execution for just moderate test sets, this can be viewed as a less exact framework due to little data set. S. Sharma and S. Gupta presented a framework where trigram and Bayesian strategies are applied to determine genuine word mistakes [5]. Not at all like different strategies that utilization 2 or 3 highlights, their proposed framework utilizes every one of the elements of the sentences. Md. M. Haque et al. planned a stochastic language model where they utilized unigram, bigram, trigram, ease off and erased addition for single word forecast [6]. Md. M. Rana et al. shown a strategy utilizing bigram, trigram and Markov presumption to discover and fix homophone blunder in genuine word mistake in Bangla language [7]. By utilizing bigram and trigram blend and removing setting highlights they recognized the blunder and created choice rundown against the competitor word dependent on likelihood computations. Notwithstanding the continuous events of bigram, trigram is given main goal in their model as it extricates more elements about the specific circumstance. A N-gram and semantic gram approach-based framework was presented by K. Wiegand and R. Patel to foresee non-syntactic word for augmentative and elective correspondence [8]. They assessed and showed the execution of four calculations. A. Jain and M. Jain proposed a model dependent on n-gram and word reference query strategy for Hindi non-word spelling blunders [9]. They showed that accessible techniques are not appropriate for applying in Hindi language which is like the proposed case. M. F. Mridha et al. zeroed in on distinguishing unexpectedly missed words while composing a sentence utilizing bigram and gave ideas utilizing trigram [10]. Accepting that bigram's premier and endmost word to be trigram's preeminent and endmost word sequentially, an idea list is produced for the missing word. Their technique fizzles now and again where despite the fact that a word is missed at this point the sentence seems right. H. Stehouwer and M. van Zaanen focused on the issue of befuddling words where sets of words are comparable and regularly utilized inaccurately in setting [11]. They utilized a nonexclusive

classifier dependent on a n-gram language model to foresee the right word in setting however its precision is relied upon to be not exactly a particular methodology. O. F. Rakib et al. proposed a model that applies GRU (Gated Recurrent Unit) put together RNN approach with respect to an informational index where they utilized uni-gram, bigram, trigram, 4-gram, 5-gram informational collections [12]. Their model can propose the following probably word as well as recommends total sentences at the same time from a given word grouping in Bangla giving 78.15% exactness. P. P. Barmana and A. Boruah planned a framework that predicts the following Assamese word utilizing Long momentary memory (LSTM) with 88.20% exactness [13]. S. Islam et al. utilized the LSTM model for producing Bangla Sentence where they made their model clamor free by eliminating accentuation denotes, extra spaces, and newlines [14]. Bangla text age by using Bidirectional RNN on N-gram informational index was proposed by S. Abujar et al. which can't identify mistake [15]. They zeroed in on developing a bidirectional RNN for setting up their model. They worked with fixed-length content and couldn't make subjective length content. T. Ghosh et al. examined and showed an examination between various strategies of Bangla Handwritten Character Recognition to feature restrictions and future extents of existing techniques [16]. E. A. Santos et al. discussed a model that distinguishes and fixes sentence structure mistakes in the English language [17]. They applied n-gram models and LSTM for displaying source code to track down punctuation blunders and combine the fixes. S. Islam et al. planned a RNN model which gives an answer for three kinds of sentence adjustments like auto-finishing, wrong course of action, and missing word [18].

2.3 Comparative Analysis and Summary

In 2021 deep learning has already evolved drastically and there has been a lot of deep learning algorithms and architectures. These algorithms are specialized in certain tasks, as some do a better job at image processing some are good at sequence analysis. In general, RNN algorithms are better at processing sequential data like time-series data, the stock price for example. Text data are also sequential data as they follow a grammatical sequence of a language. An artificial recurrent neural network that solves the vanishing gradient problem is a special type of RNN cell that is named, a Long Short-Term Memory Cell or LSTM cell for short. LSTM cells have shown better results in a lot of situations than other

RNN cells. That is a reason why we choose to use the LSTM model in our research to find a solution for the Bangla real-word error problem.

2.4 Scope of the problem

There have not been many works done on this theme in the Bangla language so we can say that there are numerous degrees to update this system. This system will identify and address Bangla real-word error and each and every individual who writes in Bangla will be benefited from this system for checking the legitimacy of Bangla sentences.

Fundamentally, we zeroed in on recognizing Bangla real-word error and recommend a remedy as per the setting of the sentence.

Along these lines, researchers can make this system more viable by preparing and utilizing cutting-edge deep learning models in the future. And furthermore, any scientist can prepare this system with a better and greater corpus for improving results.

2.5 Challenges

No single research works have been done without facing challenges. Challenges are the things that make the end result more satisfactory. Gathering and handling the dataset are the primary difficulties of this work. Uniquely, we confronted bunches of new difficulties for gathering the information. We scraped bunches of news articles on various subjects to get the data set. It was hard managing the dataset. For cleaning and normalizing the dataset, we utilized distinctive sorts of steps and strategies. In the wake of preparing every one of the information with various layers with distinctive size's age took so long in our machine. In this way, keeping a decent persistence for getting the last result was required for us. There could have been no other dataset or examination connected with our review, in this manner, it was hard to oversee and go ahead to get the most elevated level of exactness in our results. Lastly, we do acknowledge that the end result can be done better.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The subject of this research is to find out the solution for Bangla real-word error. This is a research area for Natural Language Processing or NLP for short. The problem of handling real-word error is a semantic error in terms of NLP. A semantic error can occur not only in natural languages but also in programming languages too. The semantic error refers to the violation of the rules of the meaning of a natural language, or programming language in that manner. The instrument or tools we used to do the research was very standard tools for any deep learning research. The operating system for our computer was a Windows10 operating system, Our environment for the work includes TensorFlow, NumPy, Matplotlib, Pandas, and some other Python libraries with a python 3 virtual environment.

Most of our research work was done with a Jupyter Notebook rather than a Text Editor or other Python IDEs. All of this work can be also done in Google Colaboratory.

3.2 Data Collection Procedure/Dataset Utilized

The data used in this project are Bangla news article data, which was collected through scrapping newspaper websites. All the data are from the same newspaper. The dataset contains 1500 news articles 1000 of them were used to train the LSTM model and 500 of them were processed to be used in testing the model.

3.3 Statistical Analysis

There was a total of 27176 words in the training corpus. And some of them occurred only once or less than 10 times. So, we discarded rarely occurred words and only took 10,000 words to work with.

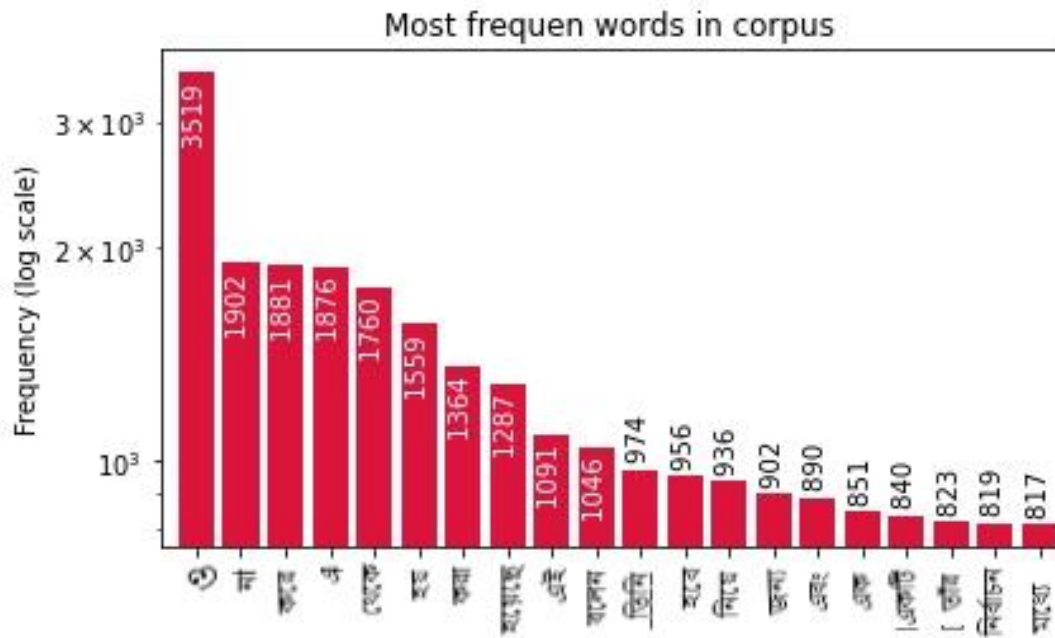


Figure 3.1.1: Most Frequent Words in Corpus

Figure 1 shows most frequent 20 words in the corpus. Unfortunately, matplotlib doesn't support Bangla language yet. So, I will list the words below with their frequency.

1. 'ও': 3519
2. 'না': 1902
3. 'করে': 1881
4. 'এ': 1876
5. 'থেকে': 1760
6. 'হয়': 1559

7. 'করা': 1364
8. 'হয়েছে': 1287
9. 'এই': 1091
10. 'বলেন': 1046
11. 'তিনি': 974
12. 'হবে': 956
13. 'নিয়ে': 936
14. 'জন্য': 902
15. 'এবং': 890
16. 'এক': 851
17. 'একটি': 840
18. 'তাঁর': 823
19. 'নির্বাচন': 819
20. 'মধ্যে': 817

It is important to know the words in the corpus to build an effective model. As some of the words in the corpus is Noun and occurred very rarely as some might just occurred in a particular article only.

3.4 Proposed Methodology/Applied Mechanism

3.4.1 Trigram Sequence Generation and Pre-padding

Trigram sequence generation is the process where every sentence is divided into three-word sequence in a way that every word in the sentence gets paired wThe sequence starts with a zero for the first two words in a sentencece starts with a zero. An example will help understand it better, as “বৈশাখে আম পাকে” will be sequenced as “0 বৈশাখে আম”, and “বৈশাখে আম পাকে”. The reason for that is, when we make our training data, the first two word will be our feeding data, and the last word will be our label data. Such as, for “0

বৈশাখে” we get “আম” as our label. Similarly, for “বৈশাখে আম” we get “পাকে” as the label. Making the dataset this way helps the LSTM model get some context of the sentence and can then identify words that does not match the context of the sentence. Text processing and training data creation process is shown in Figure 3.1.2.

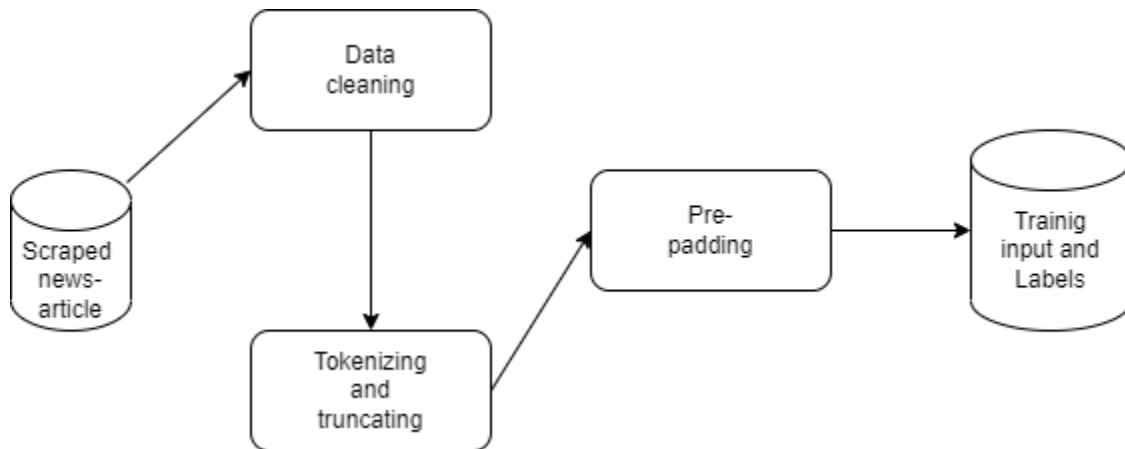


Figure 3.1.2: Text Processing and training dataset

3.4.2 Bidirectional LSTM Approach

Bidirectional Long Short-Term Memory (LSTM) networks are an extension of traditional LSTM that have the sequence information in both directions forward and backward. This kind of Recurrent Neural Network were designed to avoid long term dependencies problem. Bidirectional LSTM allows the model to recognize a sequence better by feeding it an input sequence into both directions, from beginning to the end and from end to beginning. There may arise some cases where we need more context to grasp a text as the beginning of the sentence might have an influence on the last word of a passage. Identical to this example, "ফুটবল একটি জনপ্রিয় খেলা" here, engendering the third word might be

troubling as the word can be the both "ফুটবল" as well as "ক্রিকেট". Additionally following into the context using LSTM can help understand the dependencies of generated sequences and choose the correct word. In this project, a sequential model of Bidirectional LSTM was used which consists of an Embedding layer, two Bidirectional LSTM layer of 1000 neurons and relu activation function. Following these layers, was a dense layer of 1000 neurons and last one is a dense layer consisting of neuron equal to 10,000 as of most frequent words in the training set and softmax activation function. The model was trained for 75 epochs with a batch size of 64. The neural network model uses a dense layer as the last layer which uses softmax activation function. Softmax activation function uses probability-based prediction function, it is a mathematical function that changes over a vector of numbers into a vector of probabilities, where the probabilities of each value are corresponding to the overall size of each value in the vector. So, by using softmax activation in our last dense layer of the network we can predict the likelihood of a word occurring after a sequence of two word out of the 10,000 most frequent words. By crisscrossing a whole sentence, the neural network that have been trained will predict the correction of an error word in a sentence. It will automatically distinguish whether the word should be substituted or not, hence handling both error detection and correction of the word with most likely correct word with the help of formerly fed training data-set. The neural network architecture is in figure 3.2.1.

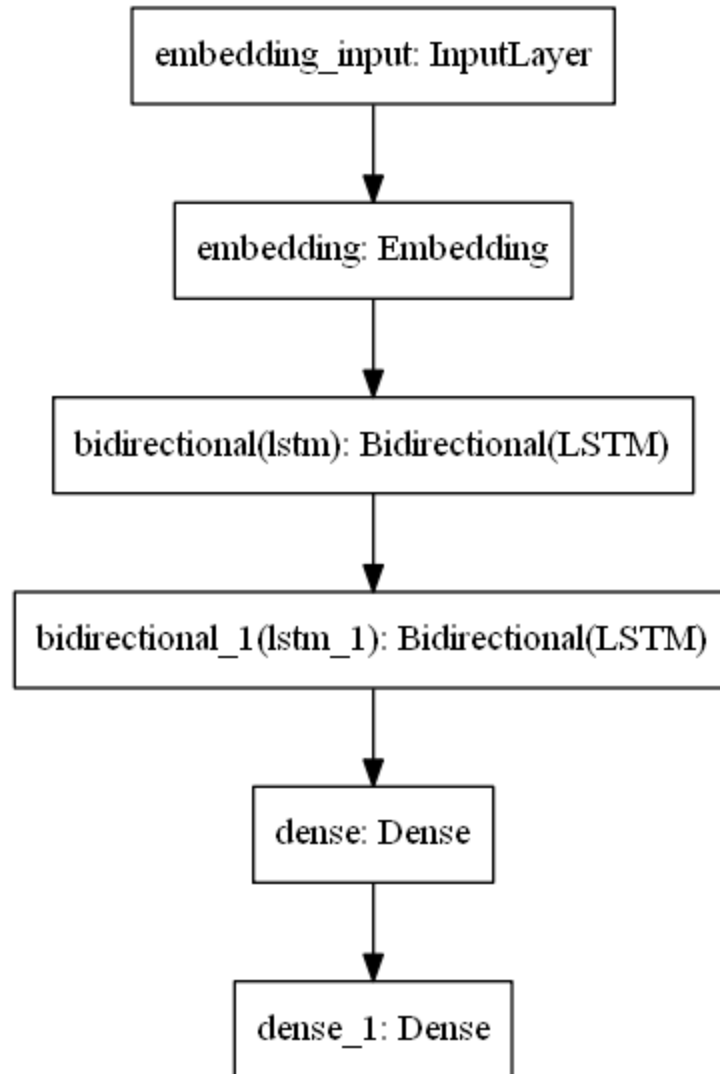


Figure 3.2.1: Neural Network Architecture

The workflow of the system is demonstrated in Figure 3.2.2.

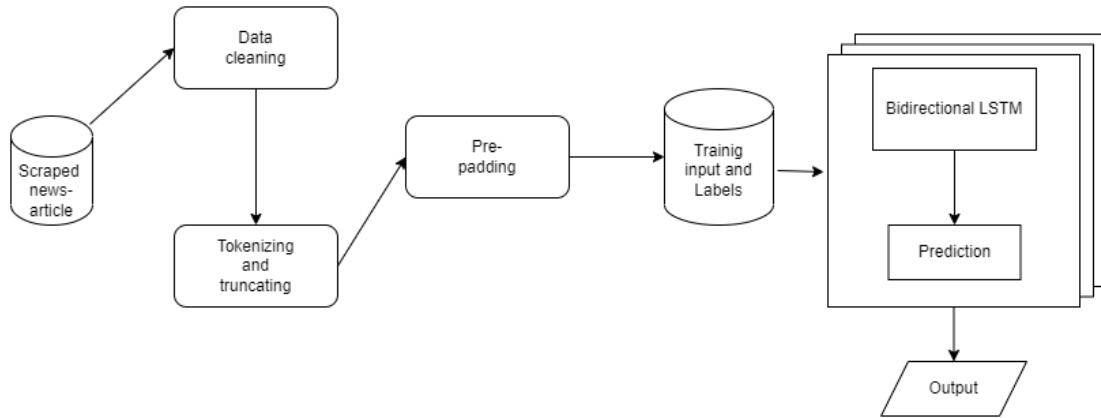


Figure 3.2.2: System Architecture for Bangla Real Word Error Detection

3.5 Implementation Requirements

1. Python 3.8 +
2. Numpy 1.3 +
3. Pandas 1.3.4 +
4. Tensorflow 2.3 +
5. Regular Expresion/ python re

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

For the experimental setup, we used an Anaconda environment where all the required python libraries were installed. For this experiment, we needed some python library for various purposes, such as computation, visualization, text processing, and other needs.

For the testing of the implemented model, we prepared 102 test sentences, where 47% of the sentences were grammatically wrong, with real-word error. The model successfully recognized the error 74.450% of the time.

4.2 Experimental Results & Analysis

The model could predict the next word from the context of a sentence with 85.47% accuracy. Which is good, but could be better with a more powerful computational machine where we could train a corpus with more article and more words. But unfortunately, currently we do not have access to such computational machine.

Table 2: Accuracy

Training Accuracy	Testing Accuracy
85.47%	74.450%

As it stands, the model is very news article focused and also, it's another limitation was the dataset we could use to train the model. In spite of that the model worked really well on news article similar to what it was trained on.

Table 3: Input/Output Assessment

Example	Input	Output
1	নিহত ব্যক্তিদের অধিকার ছিলেন শ্রমিক	নিহত ব্যক্তিদের অধিকাংশই ছিলেন শ্রমিক
2	সংকটে মানুষ কার শ্রেষ্ঠতা নিয়ে দাঁড়ায়	সংকটে মানুষ তার শ্রেষ্ঠতা নিয়ে দাঁড়ায়

For the detection and correction of test sentences the model was accurate 74.450% of the times. But fails at detecting other format of sentences, such as sentences used in novel or general conversation. This limitation can be easily overcome with appropriate dataset and a more robust and bigger LSTM architecture.

4.3 Discussion

In this project we tried to tackle a semantic error problem in Bangla language. This project has been a success, although there is much to improve on. Our final test accuracy is 74.450% but our training accuracy was 85.47% which indicates overfitting on our training data. This type of behavior is very common for RNN cells, but the result could be further improved with unbiased and larger Bangla dataset with better structural sentences. Our goal was to find out if a deep learning approach can handle the problem of real word-error detection and correction. We have succeeded on our research and the method shown in this project clearly shows that a deep learning approach to solve the real word error in Bangla language is a valid approach. So, further down this road we can expect better and more robust solution to this problem with a better deep learning algorithm and improved dataset.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

The impact on society of this project is indirect and subtle. From professional writers to students, everyone can benefit from a system where unintentional and unheeded mistakes are taken care of. People don't have to spend a lot of time looking for mistakes that often goes unnoticed. Book publishers and editors will have an easier time doing their job and can be more productive. Books and other publications will have fewer errors that makes a reading experience much better. As of now, there is no established real word error correction system in Bangla language. So, the impact on society is dreadfully clear. Language is a vital part of any society and for Bangladeshi people it is something that we fought war for nine months. So, anything that enriches our beloved Bangla language is something impactful. This project is merely a step toward making a digital Bangladesh, where we can compete and work toe to toe with the evolving technology. The world today is very much technology dependent and it will continue to be so. There for we need to be up to date with the ever-changing technology and make a Bangladesh that is technologically advanced.

5.2 Impact on Environment

Impact on the Environment is not very clear cut for this project. The system we designed will run on existing computing devices, such as mobile phone, computer and tabs. With already existing electronic devices the impact on environment will not change very much. We use mobile, tablet and computers in our daily lives. It is necessary in this modern world to use devices that will aid us in our work or daily life in general. Our system if goes on production will be implemented in a software, like digital keyboards or writing software.

For training the model, we used negligible electricity as it is no more than the amount, we already use in our house daily. So, we can see the impact of this project is not necessarily very impactful or clear.

5.3 Ethical Aspects

There is no ethical dilemma of this project and it is totally ethical as well as upright. The project is about aiding a writer to write something in Bangla as smoothly as possible. The project's motivation was to create a digital system where Bangla texts were checked for a better passage or para or just a sentence that can be written. This kind of system can already be seen in the English language and a few other languages but due to the complexity of Bangla grammar and also the lack of research on this topic, it was never been completely done. So, in this expect there are not really any ethical issues with the project.

5.4 Sustainability Plan

Computer programs are extensively used for different purposes. Commercially or noncommercially any computer program should be sustainable. Due to this project being heavily Deep Learning focused the sustainability is rather good. Apart from training the model, the trained model can be run on most computers or computing devices available nowadays. Sustainable software is software that is constantly being assessed in relation to the process of green and sustainable software engineering. This project is reliable and does not compromise future generations' opportunities.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE WORK

6.1 Summary of the Study

In short, the research was a success. The results implicate that the approach we took could very well be used in detection of real word error in Bangla sentences. Not only in specialized sector of Bangla writing but in more generalized form too. The proposed method successfully detected and corrected Bangla real word error in Bangla news articles and indicates a future scope for more research on this topic. The problem of detecting real word error in a sentence was not so easy to tackle in the past, but with the development of powerful Deep Learning algorithms, it seems that this kind of problems are not impossible or even most difficult problems in NLP.

6.2 Conclusions

In this report, we have suggested the method of acquisition and repair of Bangla

real name error based on Bigram and Bidirectional LSTM. Used data set

here is a major impact on model training due to the lack of available quality resources and computational power. However, the proposed process achieved 74.450% accuracy in correcting the real word error where the word is systematically correct but change the original meaning of the sentence. Although the model was trained on more than four data sets each with an average of 20 thousand words, we look forward to further research on a variety of contexts and more balanced data in anticipation of gaining more interesting performance.

6.3 Implication for Further Study

Further research of the projects includes using a bigger quality corpus. Trying more recent and state-of-the-art Deep Learning Transformer models, such as BERT, ALBERT, RoBERTa for the detection and correction of real-word error is in the scope for future study. The result found in this research indicates that Deep Learning algorithms could be the answer where we could really see the solution for this kind of NLP problems that were previously been seen as difficult problems for Applied Computer Science.

REFERENCES

- [1] B. Kundu , S. Chakraborti, S. K. Choudhury.: NLG Approach for Bangla Grammatical Error Correction. In: 9th International Conference on Natural Language Processing Macmillan Publishers, India (2011).
- [2] K. M. A. Hasan and M. Hozaifa, S. Dutta.: Detection of Semantic Errors from Simple Bangla Sentences.In: 17th Int'l Conf. on Computer and Information Technology, Daffodil International University, Dhaka, Bangladesh (2014).
doi:10.1016/0022-2836(81)90087-5
- [3] Z. Yuan ,T. Briscoe.: Grammatical error correction using neural machine translation. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016).
doi:10.18653/v1/N16-1042
- [4] P. Samanta , B. B. Chaudhuri.: A simple real-word error detection and correction using local word bigram and trigram. In:Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (2013).
- [5] S. Sharma ,S. Gupta.: A correction model for real-word errors. In: 4th International Conference on Eco-friendly Computing and Communication Systems (2015). doi: 10.1016/j.procs.2015.10.047
- [6] Md. M. Haque, Md. T. Habib, Md. M. Rahman.: AUTOMATED WORD PREDICTION IN BANGLA LANGUAGE USING STOCHASTIC LANGUAGE MODELS.
In: International Journal in Foundations of Computer Science Technology (2015).
doi:10.5121/ijfest.2015.5607
- [7] Md. M. Rana ,Md. E. A. Khan, M. T. Sultan, Md. M. Ahmed, M. F. Mridha, Md. A. Hamid.: Detection and Correction of Real-word Errors in Bangla Language.
In: International Conference on Bangla Speech and Language Processing(ICBSLP) (2018).
doi:10.1109/ICBSLP.2018.8554502

- [8] K. Wiegand, R. Patel.: Non-Syntactic Word Prediction for AAC. In: NAACLHLT 2012 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Montreal, Canada (2012).
- [9] A. Jain, M. Jain.: Detection and Correction of Non Word Spelling Errors in Hindi Language. In: International Conference on Data Mining and Intelligent Computing (ICDMIC) (2014). doi:10.1109/ICDMIC.2014.6954235
- [10] M. F. Mridha, Md. E. A. Khan, Md. M. Rana, Md. M. Ahmed, Md. A. Hamid, M. T. Sultan.: An Approach for Detection and Correction of Missing Word in Bengali Sentence. In: International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019). doi:10.1109/ECACE.2019.8679416
- [11] H. Stehouwer, M. van Zaanen.: Language models for contextual error detection and correction. In: Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference, Athens, Greece (2009).
- [12] O. F. Rakib , S. Akter , Md A. Khan and A. K. Das , K. M. Habibullah.: Bangla Word Prediction and Sentence Completion Using GRU: An Extended Version of RNN on N-gram Language Model. In: International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh (2019).
doi:10.1109/STI47673.2019.9068063
- [13] P. P. Barmana, A. Boruah.: A RNN based Approach for next word prediction in Assamese Phonetic Transcription. In: 8th International Conference on Advances in Computing and Communication (2018). doi:10.1016/j.procs.2018.10.359
- [14] Md. S. Islam, S. S. S. Mousumi, S. Abujar, S. A. Hossain.: Sequence-to-sequence Bangla Sentence Generation with LSTM Recurrent Neural Networks. In: International Conference on Pervasive Computing Advances and Applications – PerCAA (2019).
doi:10.1016/j.procs.2019.05.026
- [15] S. Abujar , A. K. M. Masum , S. M. Mazharul Hoque Chowdhury , M. Hasan ,

- S. A. Hossain.: Bengali Text generation Using Bi-directional RNN. In: 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), At Kanpur, India (2019). doi:10.1109/ICCCNT45670.2019.8944784
- [16] T. Ghosh, Md. M. Abedin , S. M. Chowdhury , M. A. Yousuf.: A Comprehensive Review on Recognition Techniques for Bangla Handwritten Characters. In: International Conference on Bangla Speech and Language Processing (ICBSLP) (2019). doi:10.1109/ICBSLP47725.2019.202051
- [17] E. A. Santos, J. C. Campbell, D. Patel, A. Hindle, J. N. Amaral.: Syntax and Sensibility: Using Language Models to Detect and Correct Syntax Errors. In:25th IEEE International Conference onSoftware Analysis, Evolution and Reengineering, Campobasso, Italy (2018). doi:10.1109/SANER.2018.8330219
- [18] Sadidul Islam, Mst. Farhana Sarkar, Towhid Hussain, Md. Mehedi Hasan, Dewan Md Farid, Swakkhar Shatabda.: Bangla Sentence Correction Using Deep Neural Network Based Sequence to Sequence Learning. In: 21st International Conference of Computer and Information Technology (ICCIT) (2018). doi:10.1109/ICCITECHN.2018.8631974

APPENDIX

Appendix A: Project Reflection

We began to research this project from Fall-2020. we did a lot of research on this topic. We have tried to learn about NLP, NLP in Bangla language, real-word error, research on real-word error detection in Bangla, Deep Learning applications in NLP, Language processing, and Bangla corpus in NLP research. We have collected much information about the work on real-word error detection in other languages besides Bangla. Then we decided for building a model with N-gram and LSTM and started collecting information about available Bangla corpora. Then we choose the Bangla news article for the research because it is relatively easy to gather and has a strict set of language architecture. This can help news writers on their article writing if succeed.

This was hard to try to figure out how we could use Deep Learning for the solution of the problem. We then tried and failed on many attempts for the solution of this problem. So, when we finally succeeded in detecting and correcting Bangla's real-word errors, it was a very satisfactory and fulfilling experience. Our proposed method will help take this research go further from where it was before. So, we hope this research encourages more researchers to work on this topic and take this research further.

PLAGIARISM

Plagiarism Report:

