Redefining Bangla Text Processing with a New Stemming Methodology

BY

Md Istiak Tanvir
ID: 221-15-4720
AND

Asma Akter
ID: 221-15-4636

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

Supervised by

Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Co-Supervised by

Dr. MD Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH
JANUARY 2025

APPROVAL

This Project titled **"Redefining Bangla Text Processing with a New Stemming Methodology"**, submitted by Md Istiak Tanvir, ID No: **221-15-4720** and Asma Akter, ID No: **221-15-4636** 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 January, 2025**.

BOARD OF EXAMINERS

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

Mr. Saiful Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Amir Sohel — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

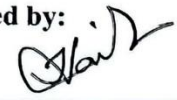
13/01/2025

Nazibur Rahman — **External Examiner**
Technical Lead - Database Administrator
Telenor - Grameen Phone Account

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Head, Department of CSE** at Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

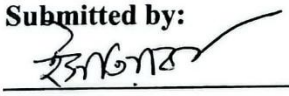


Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

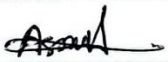
Co- Supervised by:

Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:



Md. Istiak Tanvir
Id : 221-15-4720
Department of CSE
Daffodil International University



Asma Akter
Id : 221-15-4636
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori**, **Professor & Head**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning" 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 **Dr. MD Zahid Hasan**, **Associate Professor**, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE Department of Daffodil International University.

We would like to thank our entire course mates 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 patients of our parents.

ABSTRACT

Stemming is a basic NLP technique that normalizes the linguistic structure and improves the performance of text analysis by stripping words to their roots or base forms by removing suffixes which are quite vital in applications such as text normalization, information retrieval and keeping linguistic consistency. The Bangla language has a special stemming problem due to its extensive morphological structure with many inflectional changes and intricate grammatical rules. Hence, finding the root form of a word in Bangla involves much more complex affixation patterns and compound word creation than is evident in other languages with simpler grammatical systems.

To handle the linguistic complexities of Bangla better than the previous approaches. In this work a new stemming method is developed for the language. It is more adaptable and practical for root word extraction in the real world because it has been equipped with the most updated methods to recognize and handle both the Bangla suffixes and morphological changes. Very encouraging results are obtained from the complex experiment carried out on a dataset containing 1,000 unique Bangla words. It's extremely high F1-score of 85.66% and high accuracy rate of 87.2% do, in fact, justify its superior performance over traditional stemming algorithms. It is apparent from the present study that this newly proposed stemming strategy significantly enhances the efficacy and efficiency of all Bangla text processing systems, including search engines, information retrieval platforms, and all NLP applications. This study paves the way for further development in the Bangla Language Processing problem and emphasizes the cruciality of continued research in developing language-specific Natural Language Processing tools.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Introduction

A medium that human beings use to communicate with each other is called language and Bangla is one of them. Currently the number of native Bangla speakers is 234 million and overall speakers in Bangla is 273 million are discussed by Berlitz [13]. It is a member of the Indo-European language family, same as Hindi shown in O. Sen et al [14]. In the age of the internet, Information retrieval is very important for today's world and NLP (Natural Language Processing) is the process of computational systems that helps to collect accurate information from data. So, information retrieval from those Bengali speakers from the internet is a very essential thing. Because of its complex sentence and word structure, the computer scientists came up with BNLP (Bangla Natural Language Processing) to make this language available for computers to comprehend and process by O. Sen et al [14].

One of the foundational tasks in BNLP is stemming, which is a method of reducing words to their base form which is discussed by C. Dhawan et al. [15]. For example ‘কাজটি [kajti]’, ‘কাজগুলো[kajgulo]’ , ‘কাজের[kajer]’ is the inflectional form of the word ‘কাজ[kaj]’. Here the root word is ‘কাজ[work] ‘. For this reason, stemming is essential for a variety of applications, such as search engines and text analysis where words need to be treated uniformly despite grammatical differences.

Bangla language uses its own script that derived from the ancient Brahmi script so this makes it very morphologically rich are proved by Berlitz [13]. The challenge of stemming algorithms becomes more difficult due to the wide spread, derivation and compound terms has been discussed by Das and Mitra in [16]. There is Some stemming techniques for English and other relatively simple languages that are well developed but effective

stemming solutions for Bangla are still limited has been referenced by Kazi et al. [17]. This gap presents a future task for developing a proper NLP system for Bengali language.

The objective of this work is to design and make an efficient Bangla stemmer for NLP applications which can overcome the limitations of existing methods and improve word stem accuracy in Bengali texts by integrating rule-based and our proposed algorithm. The preprocessed and clean data helps to make faster NLP analysis for collecting required Information from texts. So, one of the basic steps in NLP is to make all words to their base form for the purpose of reducing ambiguity and increasing efficiency which is known as stemming. In the context of Bangla language, there is little work in this field. It is very difficult to design a stemmer algorithm because of its complex morphological structure discussed by Salim and Hasan [1]. So, we came up with a unique and better version to solve this issue. It will also contribute to advancing the state of NLP and opening new avenues for research in Bangla language technology.

1.2 Motivation

The rapid growth of digital content in Bengali, the 5th most spoken language on the earth, has created an immense need for efficient natural language processing tools. One of the very basic subparts of NLP is stemming, which is a process of bringing down an inflection to its root form. While stemming techniques have been highly developed for some languages like English, for Bengali the progress is relatively less, since its morphology and inflectional structure are highly complicated. There are some unique challenges with regard to Bengali language processing: liberal use of suffixes, compound words and verb conjugations demand a suitable stemming algorithm. Most current stemming approaches for Bengali are either rule-based or bound to a limited dataset, which then results in inefficiency when applied to big-text analysis tasks like search engine optimization, sentiment analysis and machine translation. It is enough to recognize the challenges and motivate an investigation that will reduce the gap in computational resources for Bengali. The contribution of this work is to suggest a better stemming system that shall be helpful

in advancing the robustness of NLP applications in the Bengali language for better information retrieval, text classification, among other applications. This work also tends to empower the millions of speakers of this language by making more digital content accessible and contextually meaningful.

1.3 Research Questions

1 How can an effective stemming algorithm be developed to address the morphological complexities of the Bengali language?
2 What are the limitations of existing Bengali stemming methods, and how can they be overcome?
3 How does the proposed stemming algorithm improve the accuracy and efficiency of Bengali NLP tasks like text classification and information retrieval?

1.4 Rationale of the Study

The rapid growth of Bengali text data in online platforms, digital libraries and social media in recent times demands the development of efficient tools that can perform text processing. Though Bangla is one of the most talked languages on the earth, the NLP resources of this language are still very underdeveloped. The list includes a basic task like stemming, which has also received poor attention, thereby leaving one of the crucial gaps in the development of linguistic tools for Bengali.All these applications, from search engines to sentiment analyses to machine translations, are impeded by the lack of effective stemming algorithms for Bengali. In contrast to English or other resource-rich languages, Bengali has some unique linguistics that make conventional stemming techniques somewhat ill-equipped to achieve high accuracy. Unique features include rich sets of inflectional variants and compound word formations.

The present study is thus highly essential to overcome these limitations by proposing an efficient stemming algorithm appropriate for the Bengali language. This study will help boost the performance of downstream applications like document and information retrieval

and text classification by improving the accuracy of stemming. The implication of this study also holds far-reaching possibilities in broader attempts within language technology to ensure that speakers of the Bengali language benefit from progress being made in computational linguistics.

These findings will, therefore, fill a troublesome gap in the literature and provide the avenues of further exploration in Bengali NLP. By implication, this study will help reduce some important technological gaps and thereby provide digital empowerment to the Bengali-speaking community in making digital tools and other resources more inclusive and effective.

1.5 Research Methodology

We have developed a system that is able to identify the root forms of inflected Bengali words efficiently without complex morphological analysis. The dataset used for this system is 130k words created by merging 80k root words from Kaggle and 50k nouns from GitHub. Data cleaning has been done to remove duplicates, stop words, digits, and foreign words to assure quality.

It uses unique numeric identifiers for all 71 Bengali characters, and it assigns unique structural patterns to words for quick processing. The data is structured in a 45-indexed column by the first letter to reduce the search space and achieve improved performance; every column preserves root words along with their unique value sequence for matching, carried out in quicker form.

The system inflects input words into an integer sequence, applies inflection removal rules to that sequence, and then uses a matching algorithm to find the closest root word. If there are multiple highly similar words, it returns the shortest. Otherwise, if the original word is shorter than the best match, it returns the original word. The architecture and pseudocode show the workflow and logic of the system.

We'll go over the Experiment data, data pre-processing, data formatting, proposed algorithm implementation and algorithm evaluation. The model's perf**o**rmance will be discussed at the end of the chapter.

1.6 Expected Output

The proposed stemming system is expected to deliver the following outcomes:

1. Demonstrated improved accuracy and efficiency compared to traditional rule-based and hybrid models, particularly in processing both regular and irregular word forms.
2. Effective handling of the complexities of Bengali words for ensuring better root word identification across a diverse range of linguistic structures.
3. Recognition of areas requiring further improvement, such as handling named entity words and rare words which remain challenges for the current model.

1.7 Report Layout

This report is organized into six extensive chapters to ensure clarity and logical flow in presenting the research. Each chapter follows the preceding one in a manner that presents the study and its findings as cohesive.

CHAPTER 1: The first chapter introduces the research identifying the background and context within which the study was undertaken. It starts with 1.1 Introduction, which lays a foundational understanding of the problem statement that formed the basis of the research. 1.2 Motivation explains the motives for conducting the research and how it would add to solving existing shortcomings. 1.3 Research Questions state precisely what the research would seek to find out; this limits the scope of the research. 1.4 Rationale of the Study justifies the need for the research and how it can add value to the existing literature. 1.5 Research Methodology briefly outlines the methodologies applied in conducting the research and 1.6 Expected Output outlines what is expected in terms of output. Finally, 1.7 Report Layout encompasses a succinct overview of the report structure to guide the reader through the subsequent chapters.

CHAPTER 2: The background, basic concepts, and related works on the subject are covered in the second chapter. 2.1 Terminologies define terms that are of primary relevance

for the clarity of the work. 2.2 Related Works supports the research with literature for contextualization and identification of research gaps. 2.3 Comparative Analysis and Summary discusses the comparison in existing methodologies, thereby summarizing the positive and negative points of each. 2.4 Scope of the Problem identifies what particular problems the study covers, and 2.5 Challenges enumerates setbacks the researcher encountered during the research process.

CHAPTER 3: This chapter expounds on the methods and techniques adopted to perform the research. This covers the brief overview of methodology in 3.1 Introduction and 3.2 Data Collection Procedure, which contains the sources and how data are acquired; 3.3 Dataset Cleaning and Preprocessing, which describes how the data was prepared for analysis to ensure precision and uniformity and 3.4 Proposed Methodology, describing the novelty developed in this work. 3.5 System Architecture provides a high-level design for the system; 3.6 Algorithm elaborates on the steps involved in the proposed algorithm. Finally, 3.7 Implementation Requirements enumerates the various tools and resources utilized to implement the system.

CHAPTER 4: Chapter 4 presents the findings of the research interpretation. 4.1 Introduction: A short overview of the experimental process is given. 4.2 Experimental Setup: The technical environment and configurations used during testing have been described here. 4.3 Experiment Results and Analysis: Outcomes of the experiments are discussed here, along with the performance of the algorithm. 4.4 Discussion: In this section, the results are critically evaluated with respect to their importance and their impact on the field.

CHAPTER 5: This chapter discusses the implications of the research. Section 5.1 Impact on Society covers the way in which the findings contribute to the needs of society, especially in technology and in language processing. Section 5.2 Ethical Aspects discusses the ethical issues of the research, how this is or will be conducted responsibly. Section 5.3

Sustainability Plan outlines strategies that should be taken to ensure the longevity and adaptability of the research output.

CHAPTER 6: The last chapter synthesizes the research and sets the stage for future advances. 6.1 Overview of the Study details the important conclusions of the work and implications. 6.2 Conclusion summarizes the main points of the research objectives and its findings. 6.3 Limitations describes those areas in which the research could have been better. Finally, 6.4 Future Work represents the ideas of possible future directions of the research works that will help in encouraging progress in this field.

It provides, therefore, a clear route that the reader can go through in the research, allowing for clarity and coherence in presenting the purpose, methodology, results, and impact of the study.

CHAPTER 2
Background Study

2.1 Terminologies

1. Stemming: The process of deducting a word to its root form by removing inflections, prefixes or suffixes. For example, the word *"কথাগুলো"* can be stemed to its root *"কথা"*. It is a critical task in natural language processing for text analysis and retrieval.
2. Root Word: The simplest form of a word that cannot be further reduced and holds the core meaning. For example, in Bengali, *"করেছি,", "করবে,"* and *"করছে"* all derive from the root word *"কর"*.
3. Inflection: Variations in the form of a word to express grammatical features like tense, number, or case. For example, in Bengali, *"লেখা"* (write) changes to *"লিখছি"* (am writing) based on tense.
4. Natural Language Processing (NLP): This is a field of computer science focused on enabling machines to understand, interpret, and manipulate human language. Stemming is one of the fundamental tasks in NLP.
5. Morphology: The study of word formation and structure in a language, including how root words are modified with prefixes, suffixes, and infixes. Bengali's rich morphological structure poses challenges for stemming algorithms.
6. Algorithm: A step-by-step computational procedure designed to perform a specific task. In this study, an algorithm will be developed to accurately stem Bengali words.
7. Rule-Based Stemming: A traditional approach to stemming that relies on predefined linguistic rules, such as suffix removal, to identify root words. While effective for simpler tasks, this approach may fail to handle complex inflections in Bengali.

8. Machine Learning-Based Stemming: A modern approach where stemming algorithms learn patterns from large datasets instead of relying on manually defined rules. This can improve accuracy, especially for languages with complex grammar like Bengali.
9. Suffix: A morpheme added at the end of a root word to change its grammatical form. For example, in Bengali, *"গাছ"* (tree) becomes *"গাছের"* (of the tree) with the suffix *"-এর"*.
10. Low-Resource Language: A language with limited computational resources, such as annotated datasets, linguistic tools, and algorithms. Bengali deliberated on a low-resource language in the NLP domain.
11. Information Retrieval (IR): A process in NLP where relevant information is retrieved from a large corpus based on user queries. Accurate stemming improves IR by matching words with their root forms.

2.2 Related Work

To show its appropriate context, find gaps in research, support arguments, guide approaches, and demonstrate academic engagement, related work is important when writing a paper. It helps authors to situate their research within the knowledge corpus, ensuring that the research they carry out is opportune, original and contributes to and furthers scientific knowledge. The framework and background for the research at hand are given by related work. It helps readers understand the condition of the field's knowledge at the time and the methodologies already being applied to studies of a similar focus. Authors are able to place the findings of the study into the broader context of research and highlight their originality or uniqueness through the review of the related literature. Relevant literature reviews permit us to identify areas or gaps that require further research. Some Existing Bangla and Other Stemming research paper details given below:

1. Rule-Based Stemming: There are few works that have been proposed as a rule-based approach Such as a 'Bangla Stemmer using of rule-based approach discussed by Salim and Hasan [1]' and 'Rule-based Stemmer for Natural Language Text in

Bengali detailed by Sarkar and Bandyopadhyay [2]' which uses a set of grammatical rules and a set of fixed prefixes. But it struggles with many verb word and morphological variations because of the complex Bangla word structure. Rule-based stemming is a time expending process that is also challenging to develop and maintain. While it can be effective for certain languages but for Bangla it may not always achieve the highest accuracy rates.

2. Statistical and Machine Learning Approaches: Here are some statistical stemming techniques that rely on statistical & machine learning models such as SVM, n-grams or character-based models to identify potential stem boundaries. Some Authors are trying to solve this issue using 'the Ngram Model such as Jabbar *et al.* [3]', 'Based on contextual similarity of words like Urmi *et al.* [4]' and a 'statistical technique based on the hidden Markov models Boudchiche *et al.* [5]'. But machine learning-based stemming requires labeled data which can be challenging to obtain for low-resource languages like Bangla. Sometimes these models also may not capture the underlying linguistic rules of Bangla.

3. Hybrid Approaches and Multistep Techniques: Solving the stemming issue here also some hybrid approaches have been proposed. Hybrid approach means multi step operation that is a mix of rule based & statistical machine learning. For example, 'Hybrid Approach for Stemming in Punjabi like Dhawan *et al.* [6]', 'Stripping Affix Stemming Algorithm Using Hybrid Approach for Information Retrieval like Kadayath and Radhamani [7]'. An improved Urdu stemming algorithm for text mining based on 'multi-step hybrid approach shown in Jabbar et al. [8]'. We gather some knowledge from review papers also like [9]. These are achieving higher accuracy compared to traditional methods. However, these models face challenges in terms of scalability and efficiency.

2.3 Comparative Analysis and Summary

We reviewed several research papers that focus on Bangla stemming techniques and attempt to address this problem. For better understanding, a comparison is presented in Table 2.1.

TABLE 2.1: COMPARISON OF EXISTING STEMMING TECHNIQUE OF BANGLA LANGUAGE

Paper	Methodology	Strengths	Limitations
A Corpus Based Unsupervised Bangla Word Stemming that Using N-Gram Language Model [4]	Utilizes an unsupervised n-gram language model to generate stemmed forms by predicting the probability of word sequences in a corpus.	Effective with large datasets avoids reliance on linguistic rules.	Can overfit frequent patterns with limited contextual understanding.
Contextual Bangla Neural Stemmer: This finding Contextualized Root Word Representations [24]	Neural-based stemming using character-based neural blocks, Word2Vec and Universal Transformers to contextualize and generate root forms.	Incorporates contextual embeddings like robust for OOV words that avoids retraining BERT.	Computationally expensive but requires a pre-trained language model like BanglaBERT.
Designing a Bangla Stemmer Using Rule-Based Approach [1]	Applies pre-defined linguistic rules for suffix removal to identify root words in Bangla text.	does not require a large training corpus.	Struggles with irregular words and ambiguous rules

Design of a Rule-Based Stemmer for Natural Language Text in Bengali [2]	Focuses on a linguistic, rule-based method to remove suffixes using a stepwise algorithm and a morphological analyzer.	Structured approach for linguistic consistency is good for controlled applications.	suffers from accuracy issues with noisy or informal text.

Furthermore, while certain techniques may be transferable, they fall short of capturing the linguistic quirks of Bangla when compared to stemmers for other morphologically rich languages like Hindi, Arabic and Turkish. This emphasizes the necessity of a more effective and unique stemming solution made especially for Bangla. This literature review provides an overview of existing stemming techniques and highlights the challenges and considerations that help to fill identified gaps for designing an effective Bangla stemmer. This study aims to provide a novel approach to achieve higher accuracy and efficiency

2.4 Scope of the Problem

Despite the progress made in Bangla stemming techniques, significant gaps remain in rule-based technique or statistical machine learning technique or hybrid technique these fail to handle the complexity of irregular verbs, compound word and complex structural word. These gaps underscore the need for a more efficient and versatile stemmer for Bangla which this research seeks to address.

2.5 Challenges

The development of a robust stemming algorithm for Bengali involves several challenges due to the language's unique linguistic characteristics and technical constraints. These challenges include:

1. Complex Morphological Structure: Bengali words often consist of a root combined with multiple affixes (prefixes, suffixes, and infixes) that convey grammatical and

semantic information. This richness in morphology makes it hard to identify the exact root word. For example, *"খাবার", "খাচ্ছি,"* and *"খেয়েছে"* all derive from the root word *"খা"* but their varied structures pose challenges for stemming.

2. Lack of Annotated Datasets: A significant limitation in Bengali NLP is the scarcity of annotated datasets for stemming. Unlike resource-rich languages like English, Bengali lacks a comprehensive and diverse corpus, making it harder to train and test stemming algorithms effectively.
3. Handling Ambiguity: Bengali words can be ambiguous in their meaning or usage, depending on the context. For instance, *"জানানো"* can mean "to inform" or "to make known," creating difficulties in identifying the intended root form during stemming.
4. Suffix Overlap: Some suffixes in Bengali are common across multiple root words, leading to erroneous root identification. For example, the suffix *"-তে"* is used in both *"গাছতে"* (towards the tree) and *"খেতে"* (to eat), but their roots differ significantly.
5. Compound Words and Reduplication: Bengali frequently uses compound words (e.g., *"বইপত্র," "খাওয়াদাওয়া"*) and reduplication (e.g., *"গাছপালা," "খাইদাই"*) to convey meaning. Stemming such words accurately while preserving their semantic integrity is a non-trivial task.
6. Domain-Specific Variations: Words and expressions in Bengali can vary across domains such as news, literature and social network. These variations require the algorithm to adapt dynamically, which adds complexity to the stemming process.
7. Lack of Linguistic Standardization: Bengali has multiple regional dialects and informal usages, particularly in social media and conversational contexts. This linguistic diversity complicates the creation of rules or models that can generalize across all forms of Bengali.

CHAPTER 3
Research Methodology

3.1 Introduction

A proposed system has been developed to identify the root form of infected words in the Bengali language by leveraging the uniqueness of every single Bengali word. This system can efficiently determine the base form of a word without the need for complex morphological analysis. A depth details provided in section 3.5.

3.2 Data Collection Procedure

The dataset for this research was prepared by collecting a large dataset of Bangla words from reliable and widely used online repositories. More specifically, an 80,000-word list of root words was downloaded from Kaggle [18], a very popular platform for sharing data and resources related to machine learning. Another 50,000-word list of commonly used Bangla noun was downloaded from GitHub [19], an open-source code and data hosting service.

These two lists were joined so that a strong and diverse dataset for stemming and root word identification could be created. Joining these two lists created a rich dataset of 130,000 unique Bangla words. The dataset contains a large vocabulary representing words commonly used in day-to-day language use, enriching the system with applicability and performance.

From a large repository of heterogenous words, the proposed system is developed up to such extent that it tackles complex inflectional forms effectively to produce a superior root word for better performance of Bangla text processing.

3.3 Dataset Cleaning and Preprocessing

Data cleaning is a very significant preprocessing step for stemming to remove all the irrelevant and redundant elements present in the Bangla text corpus. In this respect, some

of the most vital tasks performed are enlisted below to enhance the quality and usability of the dataset.

1. Remove Duplicate Words: Redundant entries have been eliminated so as to avoid repetition and increase data integrity.
2. Digit-Type Word Removal: Removing words containing only digits, as they are not relevant for stemming.
3. Stop Word Removal: Removing function words that do not carry important meaning for the identification of the root word.
4. Foreign Word Exclusion: Removing words which are not Bangla in order to maintain the dataset purely in the language.
5. Empty Cell Removal: All blank cells are removed to keep the dataset clean and structured.

The listed preprocessing steps are very imperative in maximizing efficiency and accuracy of the proposed stemming system; hence, they are designed in a manner of eliminating noise to improve the quality of the input data for an efficient extraction of root words.

3.4 Proposed Methodology

Some steps are given below for better understanding of the system:

1. Understanding the uniqueness & structure of every single Bengali word -

Initially, we assign unique numerical identifiers to all Bengali characters and symbols. This allows for faster processing during calculations. Here are 71 unique numerical identifiers assigned to each of the 71 distinct Bengali characters. An example showed in Table 3.1

TABLE 3.1: SETTING UNIQUE VALUES

Symbols	‘া’	'ী’	'ি'	'ু'	……	'ড়'	'ঢ়'
Unique Value	1	2	3	4	……	70	71

When we examine the structure of a word based on the assigned unique values, we observe that each word has a distinct structural pattern. some of the examples are given Table 3.2.

TABLE 3.2: WORD STRUCTURE WITH CORRESPONDING UNIQUE VALUES

Word	Structure	List of values
মহানগর (mohanagar)	ম+হ+া+ন+গ+র (m+o+h+a+n+a+g+a+r)	[51, 59, 1, 46, 29, 54]
মামলা (case)	ম+া+ম+ল+া (c+a+s+e)	[51, 1, 51, 55, 1]
বাসা (home)	ব+া+স+া (h+o+m+e)	[49, 1, 58, 1]
সংখ্যা (number)	স+ং+খ+্+য+া (n+u+m+b+e+r)	[58, 12, 28, 13, 52, 1]

2. Understanding The datasets structure -

We downloaded 80k root words list from Kaggle [18] and 50k noun list can be found on GitHub [19] that Bengali users use in their daily life. Then we merged that and created a comprehensive dataset consisting of 130k words. After data cleaning, which included removing duplicates, null values and misspellings, we organized the dataset into 45 columns. Each column index represents words beginning with a specific Bengali letter, facilitating efficient matching of input words against this small subset rather than the entire 130k word dataset. Under each column has two sub columns, one represents the root word and another represent the unique list depend on the word latter value sequentially. It reduces the time complexity of the system and main idea for finding the root easily and accurately. An example of the dataset's column is showed in Table 3.3.

TABLE 3.3: ROOT WORD WITH CORRESPONDING UNIQUE VALUES

'অ' [index - 0]		'আ' [index - 1]		'Others'
অসভ্য (uncivilized)	[16, 58, 50, 13, 52]	আঁকন (draw)	[17, 27, 46]	…..
অনুশোচন (regret)	[16, 46, 4, 56, 9, 32, 46]	আঙটি (ring)	[17, 31, 37, 3]	…..
অগ্নিদীপক (ablaze)	[16, 29, 13, 46, 3, 44, 2, 47, 27]	আত্মলোপ (suicide)	[17, 42, 13, 51, 55, 9, 47]	…..
অনন্য (unique)	[16, 46, 46, 13, 52]	আধুনিকতা (modern)	[17, 45, 4, 46, 3, 27, 42, 1]	…..
অর্থনীতি (economy)	[16, 54, 13, 43, 46, 2, 42, 3]	আর্দ্রতা (humidity)	[17, 54, 13, 44, 13, 54, 42, 1]	…..
…….	……..	……..	……	…..
অশ্বচারণ (horse riding)	[16, 56, 13, 49, 32, 1, 54, 41]	আবির্ভূত (appeared)	[17, 49, 3, 54, 13, 50, 42]	…..

3. Understanding the workflow of the System -

The following is the flow chart of a broad view of the Bangla word root detection algorithm. It provides a systematic and logical approach that is designed in such a manner to extract the root form of Bangla words by analyzing character patterns, suffixes and similarity measures. The initiation of this process starts with the validation of input words,

whereby the algorithm checks whether the input word contains Bangla characters only. If the input fails this validation, it returns the original word otherwise it initiates the next step.

Secondly this stage deals with eliminating the noun suffix. The program determines if the word contains any of the predefined noun suffixes that are often used in Bangla. If so, the word's suffix is eliminated, and the original term is produced. If not, the suggested algorithm moves on to a more complicated step that involves matching data from a dataset or dictionary.

In the next step, the word is first numerically encoded, whereby the characters are changed to symbolic values. After encoding, the word is then worked upon to find out its root from a dataset that is indexed based on its starting character. A sequence-based search compares the input word with the indexed entries by calculating the similarity scores for finding the closest match. It picks up the word with a maximum similarity score but performs further checks to ensure the accuracy of the detected root.

If the length of the maximum similarity word is shorter or equal to two, then an additional check is performed in a data set containing words of exactly two characters. The outcome from that check determines whether to return the matched word or the original word. During the entire process, the algorithm ensures strong error handling so that if no good match is found, it returns the input word as the output.

3.5 System architecture

The system diagram has shown in Figure 3.4.

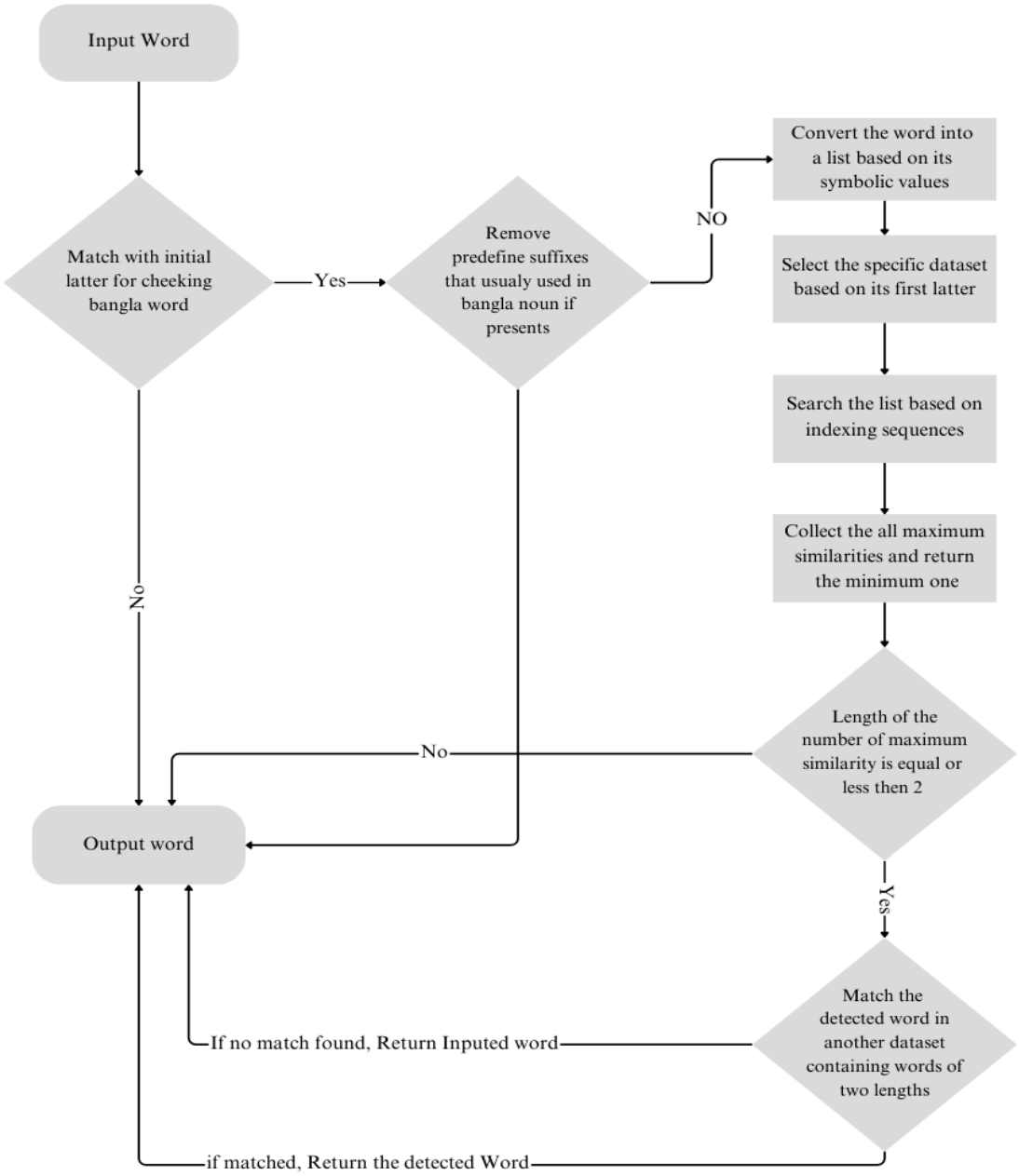


Figure 3.4: Diagram of proposed system

3.6 Pseudocode algorithm

The Proposed pseudocode algorithm are given below:

Step A: Language Validation

1. Verify if the input_word contains only Bengali characters:
    - If true: Proceed to the next step.
    - If false: Return the original word and exit the process.

Step B: Noun Suffix Check

1. Check if the input_word ends with any suffix from the predefined list of noun_suffixes:
    - If a suffix is found:
        - Remove the suffix from the word.
        - Return the modified word and exit the process.
    - If no suffix is found:
        - Proceed to the next step.

Step C: Dictionary Lookup

1. Encode the input_word numerically using character-to-integer mapping to obtain encoded_word.

2. Check if encoded_word exists in the dictionary:
    - If it exists:
        - Return the corresponding root word from the dictionary as root_word and exit the process.
    - If it does not exist:
        - Proceed to the next step.

Step D: Dataset Matching

1. Access the dataset and identify the index corresponding to the initial letter of the input_word.
2. Calculate similarity scores between the input_word and words in the dataset based on their character sequences.
3. Identify the word with the maximum similarity score.

Step E: Decision Making

1. Select the word with the highest similarity score as the potential root_word.
2. Compare the lengths of the potential root_word and the original input_word:
    - If the root word's length is greater than the input word's length:
        - Return the input_word as is and exit the process.
    - Otherwise:
        1. If the maximum similarity score is greater than two:
            - Return the detected word as the root_word and exit the process.
        2. If the maximum similarity score is two or less:
            - Match the input_word in the dataset containing words of length two.
            - If a match is found: Return the matched word and exit the process.
            - If no match is found: Return the original input_word and exit the process.

CHAPTER 4
Result Analysis and Discussion

4.1 Introduction

This paper, therefore discusses the development and evaluation of a new Bangla stemming algorithm that can overcome these problems. Our proposed system uses superior techniques in correctly identifying the root words from any Bangla text and outperforms other stemming algorithms. A wide dataset consisting of 1000 unique Bangla words representing different parts of speech and linguistic constructs was used to test the overall performance. The proposed system is compared with other Bangla stemmers and the result shows superior accuracy. Now, in the following sections, we will discuss our proposed stemming algorithm along with results and analysis to provide a broader overview of its abilities and weaknesses.

4.2 Experimental setup

In this section, we provide details about implementing the proposed stemming algorithm, developed as part of this research. The algorithm has been implemented as a Python package and made publicly available on the Python Package Index (PyPI) for easy integration into natural language processing (NLP) pipelines. we set the proposed algorithm name is bstem(BanglaStem) and package name banglanlp and its reference is in [20]. The project source code link in GitHub [21]. It can be installed and setup from PyPI using the following command:

Installation: pip install banglanlp

Setup: from banglanlp.rootfinder import bstem

Use : print (bstem(“ ঢাকায় অনেক মানুষের ভিড়”))

Output : “ঢাকা অনেক মানুষ ভিড়”

4.3 Experiment Results and Analysis

We use Accuracy and F1score formulas from [22] for calculating the performance of the systems that are commonly used in [23] as evaluation metrics. To enhance the quality of our result analysis, we have utilized three stemming algorithms: Bangla Stemmer which is available on PyPI [10], Rafi Stemmer which is available on PyPI [11] and StemmerOP which is available on PyPI [12]. The proposed system provided 87.2% accuracy and 85.66% 6 f1 score on tests at 1000 unique words which contained infected nouns, verbs, adverbs, pronouns and stop words. With the same word list other stemmer algorithms can't achieve the accuracy which our proposed system gets. The accuracy table has shown in Table 4.1 and that graphical representation shown in Figure 4.2 which demonstrate the differences between number of input word and output word and Figure 4.3 which demonstrate the comparison of accuracy and F1 Score of all Systems.

TABLE 4.1: RESULT COMPARISON

	Proposed algorithm	Bangla stemmer [10]	Rafi-kamal[11]	stemmerOP[12]
input quantity	1000	1000	1000	1000
Correct output	872	744	722	782
Accuracy (100)	87.2%	74.4%	72.2%	78.2%
F1 score (100)	85.66%	70.60%	68.43%	78.31%

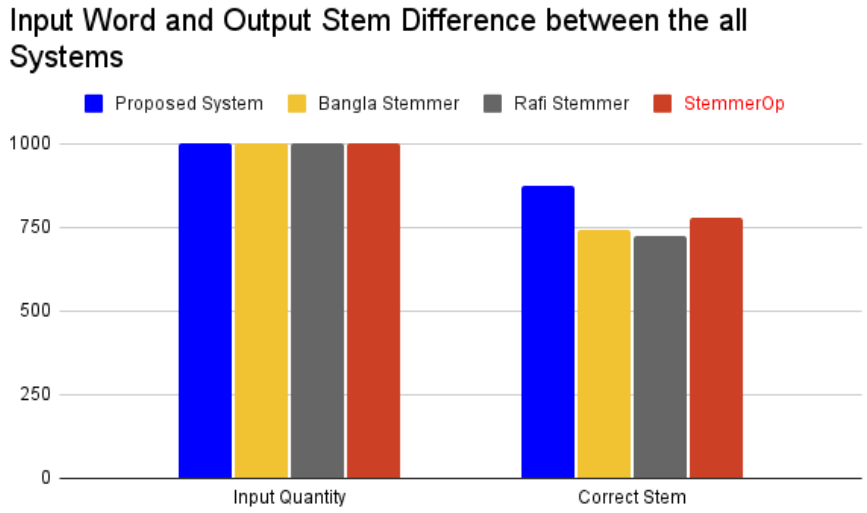


Figure 4.2: Input & Output Comparison

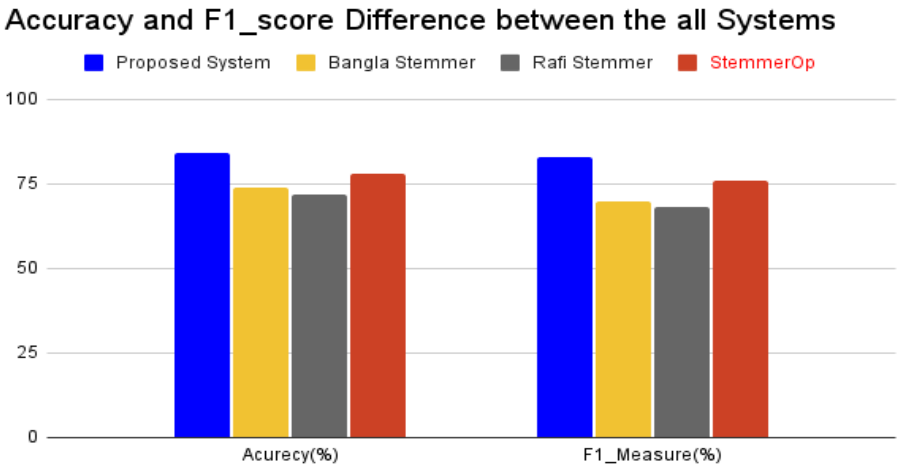


Figure 4.3: Accuracy & F1 score Comparison

And some examples of stemmed words given in Table 4.4 that are more specified in comparison with other models that struggled with certain compound words and highly irregular inflections. Despite the promising results, our proposed system also has some limitations that struggle with Name entity word and that word which does not exist in root word datasets. Its create the future scope to work on this issue. Future research could explore the long-term effects of this intervention. End of the examination, conclusion of result is that the proposed system demonstrated superior performance over traditional rule-

based models or some hybrid approaches particularly in handling both regular and irregular word forms. But further improvements are needed to address its limitations with name entity word and rare word. Future work may focus on handling the name entity word and incorporating a more sophisticated machine learning approach to enhance the model's generalization ability.

TABLE 4.4: DIFFERENCES IN OUTPUT OF ALL SYSTEMS

Input Word	Actual Root of word	Output of Proposed System	Output of Bangla Stemmer [10]	Output of rafi kamal[11]	Output of StemmerOp[12]
ব্যবহৃত (Used)	ব্যবহার (Use)	ব্যবহার	ব্যবহ	ব্যবহৃত	ব্যবহৃত
মুদ্রার (of currency)	মুদ্রা (Currency)	মুদ্রা	মুদ্	মুদ্	মুদ্রা
নেতৃত্বে (Led by)	নেতৃত্ব (Leadership)	নেতৃত্ব	নেতৃত্বে	নেতৃত্বে	নেতৃত্ব
স্থাপত্যরীতির (Architectural style)	স্থাপত্য (Architecture)	স্থাপত্য	স্থাপত্যরীতির	স্থাপত্যরীতির	স্থাপত্যরীতি
শতাব্দীর (Of century)	শতাব্দী (Century)	শতাব্দী	শতাব্দীর	শতাব্দীর	শতাব্দী

বখতিয়ার (Bakhtier)	বখতিয়ার (Bakhtier)	বখতিয়ার	বখতিয়	বখতিয়	বখতিয়

End of the examination, the conclusion of the result is that the proposed system demonstrated superior performance over traditional rule-based models or some hybrid approaches particularly in handling both regular and irregular word forms. But further improvements are needed to address its limitations with name entity word and rare word. Future work may focus on handling the name entity word and incorporating more sophisticated machine learning techniques to enhance the model's generalization capabilities.

4.4 Discussion

This system has shown clear performance leads on the standard rule-based models and hybrid approach for handling the regular and irregular word forms of the Bangla text. The result justifies the effectiveness of the combination of linguistic intuitions with a computational technique to develop higher accuracy in stemming and F1 score. The system can be further optimized for its current limitations to make the system robust for different NLP applications.

CHAPTER 5

Impact on Society, Aspects and Sustainability

5.1 Impact on Society

Developing an effective algorithm for Bengali stemming will go a long way in positively impacting society in a world where most communication, learning and information dispersal are driven by the use of digital tools. The societal level impacts are summarized below:

- Bengali Language Technology Facilitation: This, in turn, will contribute to the development of state-of-the-art language technologies such as search engines, machine translation systems and voice assistants through mitigation of the challenges in the Bengali NLP domain. Such utilities will make the language more accessible in the digital world today, hence ensuring linguistic inclusivity for millions of its speakers around the world.
- Better Information Access: A good stemming algorithm enhances the efficiency of search engines and information retrieval systems. This would facilitate easier access to relevant information by users and hence promote digital literacy, reducing the gap between technology and the Bengali-speaking public.
- Facilitating Education and Research: This work will leverage the contribution to more developed learning tools: language learning apps, e-books, and digital libraries. This will be most helpful for students and researchers in those regions where access to traditional educational resources is problematic or impossible.
- Business and E-commerce Support: The development of advanced NLP for e-commerce platforms, customer support systems and focused advertising will also help Bengali-speaking consumers and enterprises. With precise text processing in Bengali, user experiences would rise, hence growing the economy in those regions where the use of the language is relevant

- Linguistic Heritage Preservation: This would also contribute to the rich cultural and literary heritage of the Bengali language. The present study integrates Bengali seamlessly into digital platforms securing its relevance and survival in the modern era.
- Enabling Better Sentiment Analysis and Social network Monitoring: Social media platforms and organizations might be interested in the use of the Bengali content analysis stemming algorithm in sentiment analysis, opinion and trending topics. This will help in making data-driven decisions by policy makers, NGOs and businesses.
- Digital Inclusion Encouraged: Most of the Bengali-speaking rural people are still aloof from comfort regarding digital platforms because of the linguistic obstacles. The promotion of better Bengali NLP through this research will definitely help bridge such gaps and eventually result in more digital inclusion, making equal opportunities available to all.
- Encouraging Further Research: This will serve as a foundational study leading to further development in the Bengali NLP field, which will also encourage researchers and developers towards lemmatization, syntactic parsing, and sentiment analysis. It will inspire creativity and collaboration within the computational linguistics community.

The research work will fill the technological gap for Bengali and will thus create an inclusive digital society that ensures equity, accessibility, and innovation for all.

5.2 Ethical Aspects

Research in Bengali stemming and natural language processing has to consider a number of ethical issues that need to be widely addressed to make the research output socially responsible, unbiased, and inclusive. Major ethical considerations of this study are discussed herein.

- Privacy and Data Security: Any text corpus or dataset to be used for developing and testing the algorithm is sourced from publicly available sources or proper consent where required.
- Any personal or sensitive information, if any, will be anonymized so that it would not be a breach of privacy.
- Freedom from Bias: The algorithm should be designed in a manner to avoid linguistic and cultural bias and should ensure equitable representation for various dialects and regional variations of Bengali.
- Care will be taken not to foster specific word forms or structures that may place communities or dialects in a state of disadvantage.
- Transparency and Reproducibility: The approach, datasets and results should be properly documented so that other researchers may easily reproduce and verify them. This will encourage giving open access to research outputs for the benefit of the broader academic and NLP community.
- Inclusion and Accessibility: It also aims to retain the variations of Bengali in different linguistic manners, such as formal and informal usages, through the stemming algorithm. Particular emphasis will be placed on tool development for the needs of disadvantaged or digitally excluded groups.
- Responsible Use of Technology: The stemming algorithm will be designed in such a way that it can help avoid misuse with respect to spreading fake news or hate speech in the Bengali text-processing system. In this process, ethical guidelines on research will be followed to ensure that such outcomes are useful constructively for education, business and social welfare.
- Sustainability and Long-Term Impact: It also aims to provide sustainable solutions for Bengali NLP, thus fostering digital inclusion and equal opportunities in the use of technology. Any long-term impact on society and culture due to this stemming algorithm will be critically judged so that it does not cross the border of ethical and moral permissibility.

The present study focuses on the ethical issues so that the stemming algorithm may contribute towards the development of Bengali NLP by upholding some basic values related to fairness, privacy, and inclusiveness.

5.3 Sustainability Plan

A well-defined sustainability plan has been catered to in this research to ensure that the impact and usability of the outcome is long-lasting. This focuses on the sustaining relevance, accessibility and adaptability of the stemming algorithm for Bengali and its applications in NLP. Following are the main components of sustainability planning:

- Open-Source Availability
- Periodic Updates
- Community Interaction
- Integration with Emerging Technologies
- Industry-Academia Collaboration
- Documentation and Training Resources
- Monitoring and Evaluation

This will ensure that the stemming algorithm developed through this research remains a key tool for Bengali NLP and continues to serve societal benefits by fostering inclusion, innovation and cultural preservation.

CHAPTER 6

Overview of the Study, Conclusion and Future Work

6.1 Overview of the Study

This paper will discuss the design of an improved stemming algorithm in Bengali, keeping in mind the special linguistic challenges presented by its rich morphological structure. Bangla is one of the most talked languages on the earth and hence holds immense potential in digital applications. However, it is still under-represented in NLP research and its robust tools are lacking compared to other resourceful languages with respect to tasks related to stemming. After studying the existing stemming techniques and discussing their strengths and weaknesses, particularly in handling inflectional and derivational forms of words in Bengali, it then outlines deficiencies in the processing of complex morphology: compound words, irregular inflections, and regional variations by rule-based and hybrid methods. A new algorithm is therefore proposed that applies linguistic rules and computational techniques for more effective stemming. The new approach aims at handling regular and irregular word forms with more efficiency and dealing with common issues such as ambiguity, suffix overlap and compound word reduction. The research methodology also involves an extensive preprocessing pipeline where cleaning of data, tokenization and normalization have been done. Tools such as CSV and JSON were used for handling the data, and the panda's library was used for data manipulation and analysis. Standard techniques of metrics like accuracy, precision and recall are used to compare the outcome of the algorithm with all traditional models. Along with better results, limitations concerning the handling of named entity words and rare terms are also noted which may be improved with the help of deep learning techniques in the future. This research would fill the existing gap in Bengali NLP and thereby offer ground for higher-order applications such as search engines, text classification and sentiment analysis which in turn promote digital inclusiveness among the Bengali-speaking communities.

6.2 Conclusions

The development of the proposed system shows that we are successful in creating an effective Bangla stemmer that outperforms existing methods. This can enable the advancement of more accurate and robust applications such as machine translation, information analysis and sentiment retrieval. Although there are some challenges, our stemmer contributes to the advancement of NLP. Using proposed system techniques, we can think of ways to solve other NLP related tasks.

6.3 Limitations

Despite the promising results, our proposed system also has some limitations that struggle with Name entity word and that word which does not exist in root word datasets. It creates the future scope to work on this issue. Future research could explore the long-term effects of this intervention.

6.4 Future Work

Although the proposed stemming algorithm has given encouraging results concerning the handling of morphological complexities of the Bengali language, there are a few areas that need further exploration in an effort to enhance the algorithm's effectiveness and applicability. The first important area is the handling of named entity words, like proper nouns for people, places and organizations. The algorithm does not clearly distinguish between named entities and other word forms; most of these cases have resulted in inaccuracies. Future specialized modules can be created to spot named entities and leave them intact during stemming for use later on. Another critical challenge lies in addressing rare or infrequently used words that do not exist in the training dataset. Expanding the dataset to include diverse linguistic contexts and employing advanced techniques such as word embeddings or context-aware models can significantly improve the algorithm’s capability to process such words accurately. This is possible by integrating machine learning techniques, including transformer-based models or deep learning, into the

performance of the algorithm. These models acquire complex contextual and morphological patterns for better generalization and scalability when dealing with large data. Another direction of future work is the linguistic heterogeneity of the Bengali language with its rich regional dialects. The algorithm now is more or less focused on standard Bengali; regional variations remain underrepresented. Adding multidialectal support will make the algorithm more inclusive and adaptable for a wide range of linguistic contexts. Similarly, much attention should be given to improving the computational efficiency of the algorithm in order to extend its applicability to large-scale applications like big data processing and real-time NLP tasks. The use of the stemming algorithm in real-life applications, like in search engines, sentential analyses and machine translations, would serve to prove its performance in real situations. Besides, this step would underline not only the strengths of the algorithm but those that still might need further refinement. Development of the entire evaluation framework that assesses the system's performance within different linguistic and context settings will ensure that the latter is continuously enhanced and matched to the users' needs. These will extend the work at hand contributing towards the greater research in Bengali NLP and thereby contributing to technological inclusivity for millions of Bengali speakers around the world.

References

[1] M. S. Salim and K. M. Hasan, "Designing a Bangla Stemmer using rule-based approach," 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICBSLP47725.2019.201533.

[2] S. Sarkar and S. Bandyopadhyay, "Design of a Rule-based Stemmer for Natural Language Text in Bengali," in *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 2008, pp. 65-72.

[3] A. Jabbar, S. Iqbal, A. Akhunzada, and Q. Abbas, "An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 1, pp. 1–21, 2018, doi: 10.1080/0952813X.2018.1467495.

[4] T. T. Urmi, J. J. Jammy, and S. Ismail, "A corpus based unsupervised Bangla word stemming using N-gram language model," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, Dhaka, Bangladesh, 2016, pp. 824–828, doi: 10.1109/ICIEV.2016.7760117.

[5] Boudchiche, M., & Mazroui, A. (2019). A hybrid approach for Arabic lemmatization. International Journal of Speech Technology, 22(3), 563–573. https://doi.org/10.1007/s10772-018-9528-3

[6] C. Dhawan, J. Singh, and K. Garg, "Hybrid approach for stemming in Punjabi," in *Proceedings of the 2014 International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, 2014, pp. 1-5.

[7] K. S. B. Kadayath and G. Radhamani, "Affix stripping stemming algorithm using hybrid approach for information retrieval," *International Journal of Computer Applications*, vol. 10, pp. 40228-40233, 2015.

[8] A. Jabbar, S. Iqbal, A. Akhunzada, and Q. Abbas, "An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach," *Journal of Experimental & Theoretical Artificial Intelligence*, 2018, doi: 10.1080/0952813X.2018.1467495.

[9] J. Singh and V. Gupta, "A systematic review of text stemming techniques," *Artificial Intelligence Review*, vol. 48, pp. 157–217, 2017, doi: 10.1007/s10462-016-9498-2.

[10] bangla-stemmer, "Bangla Stemmer," *PyPI*, Feb. 6, 2020. [Online]. Available: https://pypi.org/project/bangla-stemmer/.

[11] Banglakit, "Bengali Stemmer," GitHub. [Online]. Available: https://github.com/banglakit/bengali-stemmer.

[12] Foysal, "sbnltk/docs/Stemmer.md at main · Foysal87/sbnltk," GitHub. [Online]. Available: https://github.com/Foysal87/sbnltk/blob/main/docs/Stemmer.md.

[13] Berlitz, "The most spoken languages in the world in 2024," Berlitz, Jul. 26, 2024. [Online]. Available: https://www.berlitz.com/blog/most-spoken-languages-world.

[14] O. Sen et al., "Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning-Based Methods," *IEEE Access*, vol. 10, pp. 38999-39044, 2022, doi: 10.1109/ACCESS.2022.3165563.

[15] C. Dhawan, J. Singh, and K. Garg, "Hybrid Approach for Stemming in Punjabi," in *Proc. Int. Conf. Computational Intelligence and Communication Networks (CICN)*, 2014, pp. 1-5.

[16] S. Das and P. Mitra, "A rule-based approach of stemming for inflectional and derivational words in Bengali," in *IEEE Technology Students' Symposium*, Kharagpur, India, 2011, pp. 134-136, doi: 10.1109/TECHSYM.2011.5783841.

[17] Kazi, K., Wohiduzzaman, K., & Ismail, S. (2018). Bangla Root Word Corpus.

[18] 80k Bangla Words List (Bangla Dictionary). (2022, December 16). Kaggle. https://www.kaggle.com/datasets/mahadivai/spelling-checker-v1

[19] Foysal. (n.d.-a). Bangla-NLP-Dataset/Bangla NER Dataset/ner_static_data_1.txt at main · Foysal87/Bangla-NLP-Dataset. GitHub. https://github.com/Foysal87/Bangla-NLP-Dataset/blob/main/Bangla%20NER%20Dataset/ner_static_data_1.txt

[20] banglanlp. (n.d.). PyPI. https://pypi.org/project/banglanlp/

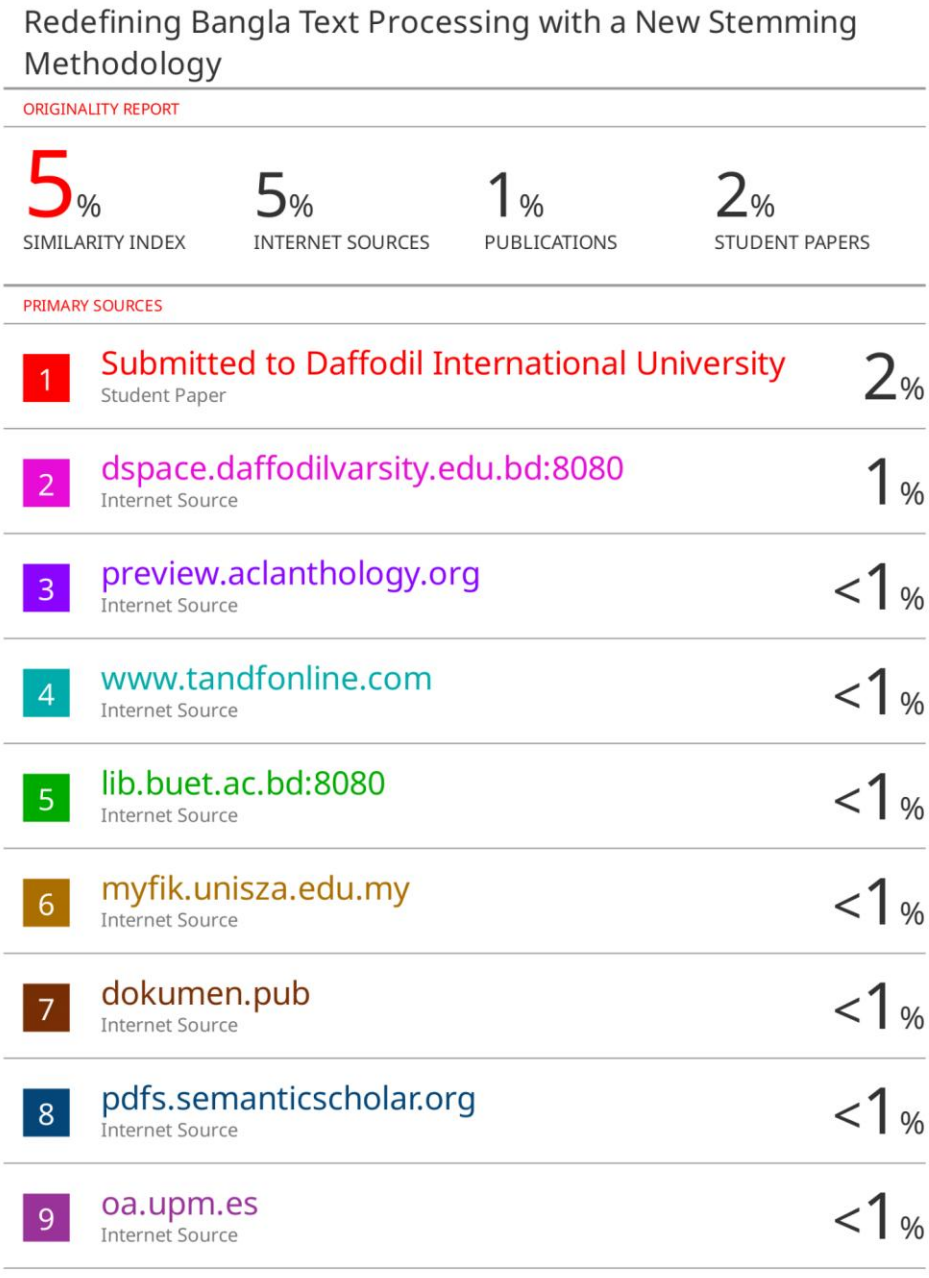
[21] ITR-Ruddro. (n.d.). Bangla-NLP. GitHub. Retrieved from https://github.com/ITR-Ruddro/Bangla-NLP

[22] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In Lecture Notes in Computer Science (Vol. 4304, pp. 1015–1021). Springer Berlin Heidelberg. https://doi.org/10.1007/11941439_114

[23] Salehin, K., Alam, M. K., Nabi, M. A., Ahmed, F., & Ashraf, F. B. (2021). A Comparative Study of Different Text Classification Approaches for Bangla News Classification. 2021 24th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 1–6. https://doi.org/10.1109/ICCIT54785.2021.9689843

[24] Md Fahim, A. A. Ali, M. A. Amin, and A. Rahman, "Contextual Bangla Neural Stemmer: Finding Contextualized Root Word Representations for Bangla Words," in Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), Singapore, Dec. 2023, pp. 94–103. Association for Computational Linguistics. Available: https://aclanthology.org/2023.banglalp-1.11.

Plagiarism Report



Redefining Bangla Text Processing with a New Stemming Methodology

ORIGINALITY REPORT

5%	5%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	preview.aclanthology.org Internet Source	<1%
4	www.tandfonline.com Internet Source	<1%
5	lib.buet.ac.bd:8080 Internet Source	<1%
6	myfik.unisza.edu.my Internet Source	<1%
7	dokumen.pub Internet Source	<1%
8	pdfs.semanticscholar.org Internet Source	<1%
9	oa.upm.es Internet Source	<1%