SUFFIX BASED AUTOMATED PARTS OF SPEECH TAGGING FOR BANGLA LANGUAGE

BY

MONJOY KUMAR ROY

ID: 151-15-5367

AND

PINTO KUMAR PAUL

ID: 151-15-5100

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

Associate Professor and Associate Head

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2018

APPROVAL

This Project titled **"Suffix Based Automated Parts of Speech Tagging for Bangla Language"**, submitted by MONJOY KUMAR ROY, ID No: 151-15-5367 and PINTO KUMAR PAUL, ID No: 151-15-5100 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 11th December 2018.

BOARD OF EXAMINARS

____

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

Narayan Ranjan Chakraborty **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

Md. Tarek Habib **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

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

DECLARATION

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

Supervised by:

____

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

Submitted by:

____

Monjoy Kumar Roy
ID: 151-15-5367
Department of CSE
Daffodil International University

____

Pinto Kumar Paul
ID: 151-15-5100
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty GOD for his divine blessing makes us possible to complete the final year project successfully.

We are really grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori**, **Associate Professor and Associate Head**, 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 draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to the Almighty GOD and **Prof. Dr. Syed Akhter Hossain, Head,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

Natural language processing (NLP) is the technique by which we process the human language with the computer. Parts-of-Speech (POS) tagging is one of the fundamental requirements for some NLP applications. It is considered as a solved problem for some foreign languages, such as English, Chinese, due to higher accuracy (97%), where it is still an unsolved problem for Bangla because of its ambiguity. Although making a POS tagger for Bangla is not a new work, but each one of available POS taggers has different kinds of limitations. We choose to develop an unsupervised system rather than a supervised system, because a supervised system needs a huge data resource for training purpose and available resources in Bangla is really poor. Here we develop a POS tagger mainly based on Bangla grammar especially suffixes. Because Bangla is a very inflectional language, where a single word has many variants based on their suffixes. In this POS tagger, we assign 8 base POS tags, where some rules, based on Bangla grammar and suffix, are applied to identify POS tags with the cooperation of verb root dataset. To handle non-suffix words, a dataset of almost 14500 Bangla words, with having their default POS tags, is added with the system, which helps to increase the efficiency of this POS tagger. A modified version of previously used algorithm for suffix analysis is applied, which result in a satisfactory level of about 94.2%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Natural Language Processing (NLP) is one of the most well-known fields of Artificial Intelligence (AI) that allows computers to process and manipulate human languages. Parts of Speech (POS) tagger is a pre-requisite part of many applications of NLP. Automated Parts of Speech tagging is the process of identifying a word in a sentence with an accurate parts of speech tag. It will contribute to almost every branch of NLP, because it has a multiform utilization to translate from one language to another language [1], to verify text data, to correct the grammatical error, to annotate the named entity in large corpus, and so on. POS tagging is also used to build Natural Language Interface (NLI) [2]. Bangla is, an Indo-Aryan language, the seventh most spoken native language; mostly used in Bangladesh and some parts of India, which has many loanwords from European languages and about 30 percent-unmodified Sanskrit words. Some of those words are modified with times and changed their actual meaning that makes Bangla words more complex and more ambiguous. Ambiguous words are more difficult to tag with a specific part of speech, because there may have different tags for a single word based on their role in the sentence. POS tagging for Bangla language is more difficult and challenging than other languages because of its prefix and suffix. Though some researchers improve a lot in POS tagging and use different methods and algorithms, there are still lots of chances to improve more.

1.2 Motivation

The people of current decade experiences the increasing number of uses of intelligent gadgets, such as amazon echo, which process human language and perform the work asked for. On the other hand, people are more active in their social media network nowadays. Sentiment can be gathered by analyzing one’s status and comments. Moreover, a computer can translate from one language to another language and assist a person to find out the best route. Most of the application is available in foreign languages like English, Japanese, Italian etc. If those applications are available in

Bangla language, that will be more helpful for us. POS tagging plays a vital role to develop those applications. Other languages like, English, Chinese etc. have their own well-developed POS tagger [3]. Some of them have over 97 percent accuracy. Bangla is the seventh most spoken native language. There are also some works on the Bangla POS tagger. Some of them are focused on verb or Noun [4] or classifications of base tags [5]. But the satisfactory level of those papers is not as good as English or Chinese language. So, we want to develop a POS tagger based on suffix analysis in order to enrich the language in the area of NLP.

1.3 Research Question

What is the effect of an unsupervised system to improve the accuracy level of POS tagging rather than the supervised system?

The purpose of this study is to find out the performance level of an unsupervised system in the area of POS tagging. Supervised system performs well in this field, where unsupervised system still have to improve. This study enables us to develop a POS tagger with a deep analysis of suffix.

Table I: Eight Parts-of-Speech tags with their description

	Tags	Description
1	NN	Noun
2	PRO	Pronoun
3	ADJ	Adjective
4	VRB	Verb
5	ADV	Adverb
6	PRE	Preposition
7	CON	Conjunction
8	INT	Interjection

1.4 Expected Output

The input to this system will be a string of words in Bangla. The POS tagger will tag each word of the input to their particular parts of speech with a developed algorithm.

The accuracy level of this unsupervised system to identify word with accurate parts of speech will be better than other available Bangla POS tagger. An application will also developed to implement the developed algorithm. In this system, we consider eight base POS tags (shown in Table I).

1.5 Report Layout

We divide this report into five section. This is the first section where we talk about motivation for our work and the expected outcome. In the second section (CHAPTER 2) we discuss about related works in this field, scope of the problems, challenges etc. In the third section (CHAPTER 3) we discuss about data collection procedure and implementation. Section four (CHAPTER 4) is for experimental result and analysis. Conclusion and future work are discussed in CHAPTER 5.

CHAPTER 2
BACKGROUND

2.1 Introduction

By using the linguistic rule or stochastic rule or both, many POS tagger was developed for different languages. Hidden Markov Model (HMM) is popular among stochastic application. But it needs large label data for better performance. And Bangla have not such large annotated dataset. Support Vector Machine (SVM), Conditional Random Field (CRF) and Maximum Entropy (ME) are also used for this purpose. Here we discuss some existing Bangla POS taggers mentioning their performance.

2.2 Related Works

At the early stage, A. Ekbal et al. [6] proposed a Bengali POS tagger using SVM, where they used a corpus of 72,341 wordforms tagged, i.e. 15K wordforms as development set and 57,341 wordforms as training set. It came out with an accuracy of 86.84%, which was much better than the existing system based on HMM [7] and ME [8].

S. Mukherjee et al. [9] compared a Global Linear Model (GLM) based POS tagger with CRF, SVM, HMM and ME based Bengali POS tagger, where GLM won by giving 93.12% accuracy. They used a training dataset containing manually annotated 44K words and two test sets containing 14,784 and 10,273 words respectably.

S. Dandapat et al. [10] developed a Bangla POS tagger using supervised and semi-supervised bi-gram HMM and a ME based model with morphological analysis. For their work, they took almost 40,000 words as training data and 5,000 words as test data. Although the result was at satisfactory level, but there has a lot of scope of improvement.

All of these work mainly based on training data, indicates supervised tagger. H. Ali [11] tried to make some difference. He did some experimental job of making an unsupervised POS tagger for Bangla language using Baum-Welch algorithm, although the experiment did not get success.

A. Parikh [12] developed a Parts-of-speech tagging system based on neural network. The tagger was tested as single-neuro tagger and multi-neuro tagger differently. Multi-neuro tagger performed better with accuracy level of testing data 92.19% than single-neuro tagger.

S. Ismail et al. [13] developed an algorithm for making an automated Bangla POS tagged dictionary. The algorithm was dedicated to tag words in Noun, Verb and Adjective using suffix list.

Md. N. Hoque et al. [14] developed a POS tagger system by applying Bangla stemmer and rule based analyzer. Some rules were generated from suffix and some were from observations. They got the accuracy level of 93.7%, but still they have many drawbacks.

2.3 Research Summary

A supervised tagging system with HMM, ME, CRF, SVM performs satisfactorily for tagging Bangla words. But due to the huge dataset requirement, it will be very laborious to develop. Its performance level depends on the size of training data. Moreover, the neural network has done this job pretty well, but still hard to develop. Now it is the term for an unsupervised system. A limited work has done in this section for tagging purposes, which performs fairly well. But all of them have a number of limitations, those make a scope for improvement. So, we develop an unsupervised tagging system, powered by an modified version of previously used algorithm [14]. In this system, each word is inspected for appropriate POS tag based on suffix analysis. There are also some rules that help to make decisions to get an appropriate tag. We are considering 8 tags (shown in Table I) for this system.

2.4 Scope of the problem

The problem is the part of an experiment. There have a number of scopes for occurring problems.

- **Manually tags dataset words**. A data set contains lots of words or sentences. Tagging each and every word with accurate POS tag is a lengthy and difficult process.

- **Crop the word into accurate root word**. The ending of a word can matches with multiple suffixes, which results in multiple roots. It is difficult to find the right one.
- **Tag multiple meaning words.** A word can be used for multiple purposes. So, there can be more than one POS tags available for that word.
- **Handle unknown word**. Bangla is a compound language, which contains many foreign words. Moreover, every day many new words enter in Bangla language. Categorize those unknown new words is difficult.

2.5 Challenges

Bangla language has a huge range of vocabulary. The number of words is increasing every day. A huge collection of words is not easily available in this language. But for the system, we need a collection of verb roots as much as possible. Resources for this language is really poor. Moreover, Bangla words are polymorphous. The form of words changes over time, which increase the difficulties of collecting the verb roots. Root, which works for সাধু form, doesn't work for চলিত form of a word.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Methodology

The goal of this study is to find out the performance of an unsupervised system in the purpose of POS tagging. To successfully conduct the thesis below steps were taken.

- The necessity of a POS tagger in Bangla NLP applications was observed.
- A study on the performance of different algorithms for POS tagging on different languages was done.
- Various papers on Bangla POS tagger were studied and a comparison between the performance of English POS tagger and Bangla POS tagger was made.
- The poor resource availability in Bangla language was realized and a decision of making an unsupervised system was made.
- As Bangla is an inflectional language, different Bangla grammar books were studied to identify the effects of Bangla inflections in Bangla words.
- Available papers on suffix analysis were shortlisted and studied.
- A collection of Bangla suffixes according to their classifications from different resources and a set of rules for identify tags with suffix were made.
- The necessity of Verb root to identify tags was realized and a collection of Verb root was made.
- A modification of previously used algorithm in POS tagging was done.
- As soon as the necessity to handle non-suffix words raised, we made a collection of words with their default tags.
- Side by side, the web application to implement this system was developed.
- Finally, an unsupervised Bangla POS tagger based on suffix analysis is proposed to increase the accuracy level.

3.2 Data collection Procedure

Our collection of data divided into two parts, one part (almost 14,500 words) used as a dictionary where each word has its manually given tag [17] and another part (almost

12,000 words) used as a testing dataset. There is another collection of data, which contains verb roots. Collection of data goes through a lengthy process. Dictionary data are collected from different kind of magazines, short stories, and novels. And Testing data are from different popular online Daily newspaper and blogs. After collecting dictionary data, each word has given a POS tag manually based on the grammatical rules [15][16] and Bengali to English dictionary [17]. On the other hand, Verb roots are gathered from Bangla grammar books [15] [16] and different online sources.

3.3 Implementation Details

Bangla is very inflectional language, where each word may have more than one meaning based on inflection. Suffix or postfix is an inflection, which does not have any meaning as itself. It sits at the end of a word or a bunch of characters to make a meaningful word. Bangla language has rich grammatical support. The grammar defines how suffixes stand with the word to make different meaning, also different parts of speech. We generate some rules according to Bangla grammar and sentence pattern to analyze the data. Those rules are mentioned below.

Rule 1: According to Bangla grammar, a Verb word is a combination of a verb root called ধাতু and an inflection known as ক্রিয়া-বিভক্তি (some of them shown in Table II), which only appears after Verb root. Suppose, current word i.e. $word_i$ (i = 0, 1, 2,…, length of sentence - 1) is examined and found with ক্রিয়া-বিভক্তি at the end. After chopping ক্রিয়া-বিভক্তি, if the remaining is found as a Verb root, then $word_i$ will be considered as a Verb. Example -

করেছিলাম = কর্ + এছিলাম

ধাতু ক্রিয়া-বিভক্তি

Table II: List of Verb suffix or ক্রিয়া-বিভক্তি

Description	List
ক্রিয়া-বিভক্তি or Verb suffix	ছেন, েছে, েছেন, িয়াছ, েছ, িয়াছিস, এছিলাম, েছিস, িয়াছি, েছি, ুন, িলেন, লাম, িতেছিলেন, েছিলাম, াচ্ছে, াইতেছিস, াচ্ছিস, িয়েছেন, াক, লো, াইলেন, াইত, াত, াতাম, াচ্ছিল, াচ্ছিলে, াচ্ছিলি, াচ্ছিলাম, চ্ছিলাম

Table III: List of কৃৎ-প্রত্যয়

Parts of Speech	List
Noun	ওন, ানো, না, ওনা, ুনি, ন, ানি, ারি, ারী, ুরি, না, ানোর
Adjective	ন্ত, ুন্তি, ুক, ুকা, কো, োয়া, িষ্ণু, বর, মান

Rule 2: There are two types of suffixes in Bangla grammar, one called কৃৎ-প্রত্যয় and another called তদ্ধিত-প্রত্যয়. Between two of them, কৃৎ-প্রত্যয় only appears at the end of Verb root and তদ্ধিত-প্রত্যয় appears at the end of a meaningful word. Example -

চলন = চল্ (ধাতু) + অন (কৃৎ-প্রত্যয়)
বাড়িওয়ালা = বাড়ি (শব্দ) + ওয়ালা (তদ্ধিত-প্রত্যয়)

Here "চল্" is a verb root and "বাড়ি" is a word. Now, the interesting thing is, কৃৎ-প্রত্যয় only appear at the end of verb root and generate only Noun and Adjective [16]. Some of them only generate Noun word and some generate only the Adjective. So, we categorize them into Noun suffix group and Adjective suffix group (shown in Table III). If a Noun suffix appears at the end of Verb root, then the word will be tagged as a Noun. The same thing happens for Adjective suffix.

চলন = চল্ (ধাতু) + ন (কৃৎ-প্রত্যয়) = NOUN
চলন্ত = চল্ (ধাতু) + ন্ত (কৃৎ-প্রত্যয়) = ADJECTIVE

Rule 3: Consider every number is expressed in digits, not in words. According to the dictionary [17] and Bangla grammar [16], a number can be Noun or Adjective. Suppose, current word i.e. $word_i$ is a number, if immediate next word i.e. $word_{i+1}$ is a quantifier marker (from the list Table IV), then $word_i$ is considered as an Adjective and $word_{i+1}$ is considered as a Noun. Otherwise, $word_i$ is considered as a Noun.

Sentence 1: "সে ১০ মাইল পথ পারি দিয়েছে।"

Sentence 2: "ধারা ১০ অনুসারে সে সাজা পেলো।"

In sentence 1, when "১০" is examined, the $word_{i+1}$ "মাইল" is found as a quantifier marker (from Table IV). So, "১০" will be considered as Adjective and "মাইল" is considered as Noun. But in sentence 2, "১০" will be considered as a Noun because next word is not a quantifier marker.

Rule 4: There are some suffixes, which stands with a number. We can divide them into two categories (Shown in Table V). One category expresses the number as Noun and another one express as Adjective. If the first category raised in $word_i$ then immediate next word $word_{i+1}$ considered as Noun. Because these Noun suffix categories actually used for indicates a date of a month. So, immediate next word must be a month's name.

Sentence 1: "সে ক্লাসে ১ম স্থান অধিকার করেছে।"

Sentence 2: "আজ ১লা বৈশাখ।"

In sentence 1, "১ম" indicates a position, so will be considered as Adjective. But in sentence 2, "১লা" indicates a fixed day of a month, so will be considered as Noun and $word_{i+1}$ also as Noun.

Table IV: Quantifier Marker List

Parts of speech	Quantifier Marker
Noun	টাকা, ডলার, রুপি, দিনার, ইয়েন, ইউরো, মাস, দিন, সপ্তাহ, সাল, বছর, যুগ, গজ, ফুট, ইঞ্চি, মিলিমিটার, সেন্টিমিটার, মিটার, কিলোমিটার, কেজি, লিটার, শতাংশ, অংশ, বার, শত, হাজার, লাখ, লক্ষ, কোটি, মিলিয়ন, বিলিয়ন, জন

Table V: Suffixes for number

Parts of Speech	List
Noun	লা, রা, ঠা, ই, শে
Adjective	টা, টি, খানা, খানি, টে, ম, য়, র্থ, ষ্ঠ, শ, তম, %

Rule 5: As mention before, তদ্ধিত-প্রত্যয় appears after meaningful words, but some of them can be categorized uniquely to identify a word as Noun or Adjective or Adverb. So, when $word_i$ is examined, if it ends with any of these suffixes (some of them shown in Table VI), then it will be considered as Noun/Adjective/Adverb respectively. For example, "বিপদজনক" ends with "জনক", which listed (in Table VI) as Adjective suffix. So, this word will be considered as an Adjective.

Rule 6: After stemming "ে" or "এ" from $word_i$, if it is found as Noun or Adjective, then $word_i$ will be considered as Adverb [15]. For example – if "চরমে" is stemmed with

"ে" and remaining "চরম" is found as Adjective, then "চরমে" will be considered as Adverb.

Rule 7: This rule is generated to handle words, which have more than one meaning. For example – "ও" sometimes used for addressing someone and sometimes used for connecting two sentences or words. For these kinds of word, we follow the maximizing rule. We search the word in the dictionary and gather unique tags for that word with their frequency. Finally, get the maximum frequency tag for that word.

Rule 8: Consider, "সুখ-দুঃখ", "তা-ও". Here, one indicates two different words connected with '-', and another indicates one word. We assume '-' containing word as two different words and apply other rules. After applying all rules mentioned above, if the word doesn't have any tag then combine those two words with '-', considered as one word and find into the dictionary. If it is not found in the dictionary then considered as a Noun.

Rule 9: If no above mention rule is applicable for the word, then it will be considered as Noun.

Table VI: List of তদ্ধিত প্রত্যয় for Noun, Adjective and Adverb

Parts of Speech	List
Noun	ওয়ালা, খানা, গিরি, দান, দানি, ুরিয়া, শীল, বাজি
Adjective	পানা, ভর, ভরা, মন্ত, বন্ত, খোর, জনক, মূলক, ব্যাপী, যোগ্য, কেন্দ্রিক, ঘটিত
Adverb	রূপে, ভাবে, ভাবেই, পূর্বক, মতো

We have modified an algorithm for tagging Bangla word based on above mentioned rules. This algorithm needs four parameters, testing dataset as corpus, tagged dictionary dataset as dic, verb root dataset as vrot and quantifier marker list as quantifier_list. The algorithm is given below –

```
banglaPosTaggarAlgo(corpus, dic, vrot, quantifier_list)

1. for each sentence from corpus
2.         for each word from sentence
3.                 take current word as word and next word as next_word
```

```
if (isVerb(word, vrot))
    continue
else
    tag = isNounOrAdjectiveFromKritSuffix(word, vrot)
    if (tag != false)
        tag = verb and continue
    end if
    if (isNumber(word))
        tag = tagForNumber(word, next_word, quantifier_list)
        if next word is also tagged
            increase one step for loop
            continue
        else
            continue
        end if
    end if
    if (isQuantifierMarker(word, quantifier_list))
        continue
    end if
    tag = getTagFromToddhitSuffix(word)
    if (tag != false)
        continue
    end if
    if(isInDictionary(word, dic))
        continue
    else
        if ‘ে’ present at the end
            remove ‘ে’ and check again in dictionary
            if get tag as Noun or Adjective
                continue
            end if
        end if
        roots = stemmer(word)
```

```
37.                 if (count(roots) != 0)
38.                     for each root from roots
39.                         if match found
40.                             continue
41.                         end if
42.                     end for
43.                 else
44.                     tag = Noun
45.                     continue
46.                 end if
47.             end if
48.         end if
49.     end for
50. end for
```

Let’s discuss the algorithm *banglaPosTaggarAlgo(corpus, dic, vrot, quantifier_list)* step by step. First, we take each sentence from the corpus in line 1 and explode it with space ‘ ’ in a word array. Then take each word from this word array in line 2 and check if it contains hyphen ‘-’ or not, for a purpose of getting current word and next word in line 3. If yes then explode it and take 1st part as current word and 2nd part as next word. If not, then take $word_i$ as current word and $word_{i+1}$ as next word.

Now, $word_i$ is going for verb checking in line 4, where current word is cropped with the list of ক্রিয়া-বিভক্তি (Table II) whenever a match found, and take all cropped roots into an array. If this array has any root that matches with verb root, then $word_i$ considered as a verb in line 5. If $word_i$ is not a Verb, then $word_i$ go for Noun/Adjective checking in line 7, 8, 9. Here, $word_i$ is cropped with Noun /Adjective কৃৎ suffix list (Table III) and takes roots to match with verb root. If match found, then gets corresponding tag according to suffix category.

From line 11-19, algorithm works for numerical words. If $word_i$ is a number containing word and it contains only digit, then check $word_{i+1}$ is a quantifier marker or not. If $word_{i+1}$ is a quantifier marker (Table IV), then tag $word_i$ as Adjective and $word_{i+1}$ as Noun. If $word_{i+1}$contains quantifier marker as substring, then tag $word_i$

as Adjective. If $word_{i+1}$ does not fulfill any of these two conditions, then tag $word_i$ as Noun. But if $word_i$ does not only contain digits then match the ending suffix with listed Noun/Adjective suffix for numbers (Table V). If match found, then get corresponding tag.

Now, If $word_i$ is a quantifier marker then tag $word_i$ as Noun, but if $word_i$ contains quantifier marker as substring, then also $word_i$ is considered as Noun in line 20, 21, 22. If $word_i$ does not get tag yet, then check whether the $word_i$ ends with Noun/Adjective/Adverb তদ্ধিত suffix list (Table VI) in line 23. If a match found, then get the corresponding tag.

Finally, the word goes for dictionary search from line 27-47. Algorithm search for $word_i$ in the full dictionary. An array contains unique tags found for $word_i$ with their frequency. And then calculate the maximum frequency tag for $word_i$. But if $word_i$ is not present in the dictionary, check whether it is end with “ে” or not. If yes, then chop the “ে” from $word_i$ and check the remainder into dictionary. If it results in Noun or adjective then $word_i$ is considered as Adverb. If “ে” is not present then it is cropped with other তদ্ধিত suffix and the search process in the dictionary is repeated. This time cropped one is search into the dictionary. If still not found into the dictionary then tag the $word_i$ as Noun.

Consider these following sentences as input for the system collected from corpus -

“কেন্দ্রীয় সরকার ২৫শে অক্টোবর জারি করে এই নির্দেশিকা। বিশ্বের ৭৪টি দেশের ৪৫০টি শহরের মানুষ বর্তমানে উবারের সেবা পাচ্ছে।”

After applying this algorithm -

কেন্দ্রীয় (ADJ) সরকার (NN) ২৫শে (NN) অক্টোবর (NN) জারি (NN) করে (VRB) এই (PRO) নির্দেশিকা (NN)

বিশ্বের (NN) ৭৪টি (ADJ) দেশের (NN) ৪৫০টি (ADJ) শহরের (NN) মানুষ (NN) বর্তমানে (ADV) উবারের (NN) সেবা (NN) পাচ্ছে (VRB)

3.4 Instrumental Requirement

a) Front-End Design

- HTML
- CSS

b) Back-End Design

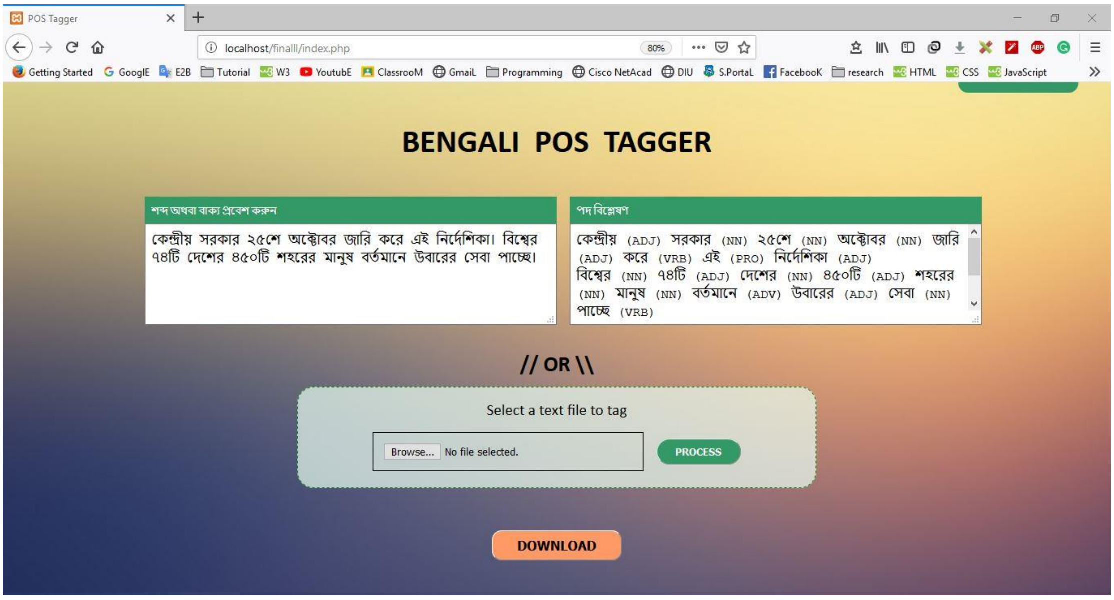
- PHP
- JavaScript
- MySQL

c) IDE Tool

- Brackets

3.5 System Design

In this section, we will present all the relevant information about our developed application, all the features it provides and how user and admin interact with the application.



কেন্দ্রীয় সরকার ২৫শে অক্টোবর জারি করে এই নির্দেশিকা। বিশ্বের ৭৪টি দেশের ৪৫০টি শহরের মানুষ বর্তমানে উবারের সেবা পাচ্ছে।

কেন্দ্রীয় (ADJ) সরকার (NN) ২৫শে (NN) অক্টোবর (NN) জারি (ADJ) করে (VRB) এই (PRO) নির্দেশিকা (ADJ)
বিশ্বের (NN) ৭৪টি (ADJ) দেশের (NN) ৪৫০টি (ADJ) শহরের (NN) মানুষ (NN) বর্তমানে (ADV) উবারের (ADJ) সেবা (NN) পাচ্ছে (VRB)

Figure 3.5.1: POS Tagger Page for User

In figure I, we represent our home page of our application, where user can input any string of words for POS tagging purpose. There are total two text area for processing the text, one for input text and another for output text having their POS tags. If the input string is large enough, then user can also input the string as a text file for processing. Every time, when system process the string, a download option appears for downloading the output result as a text file.

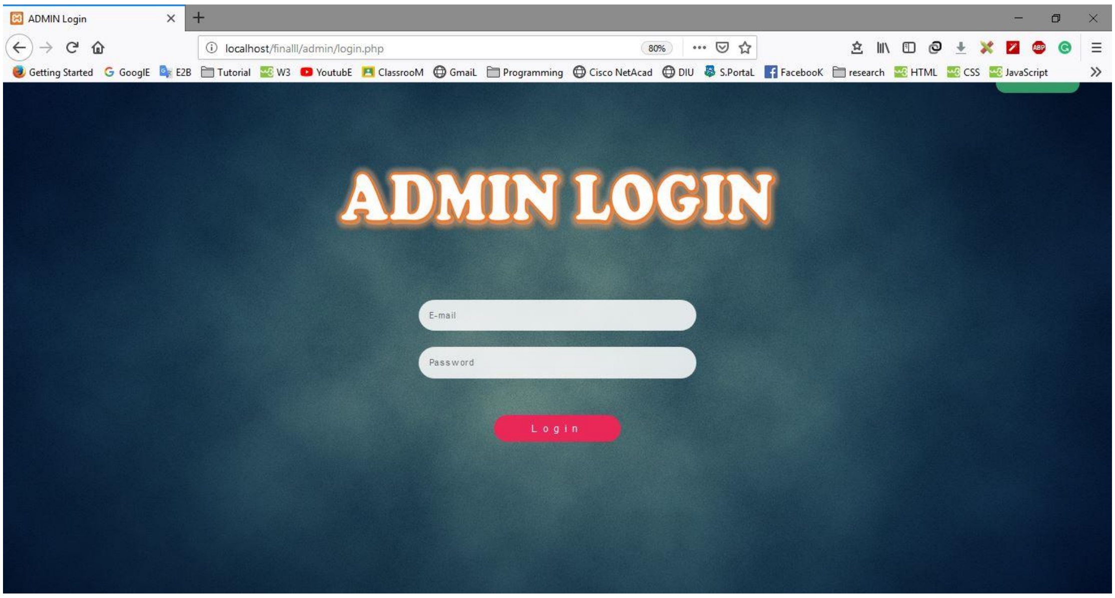
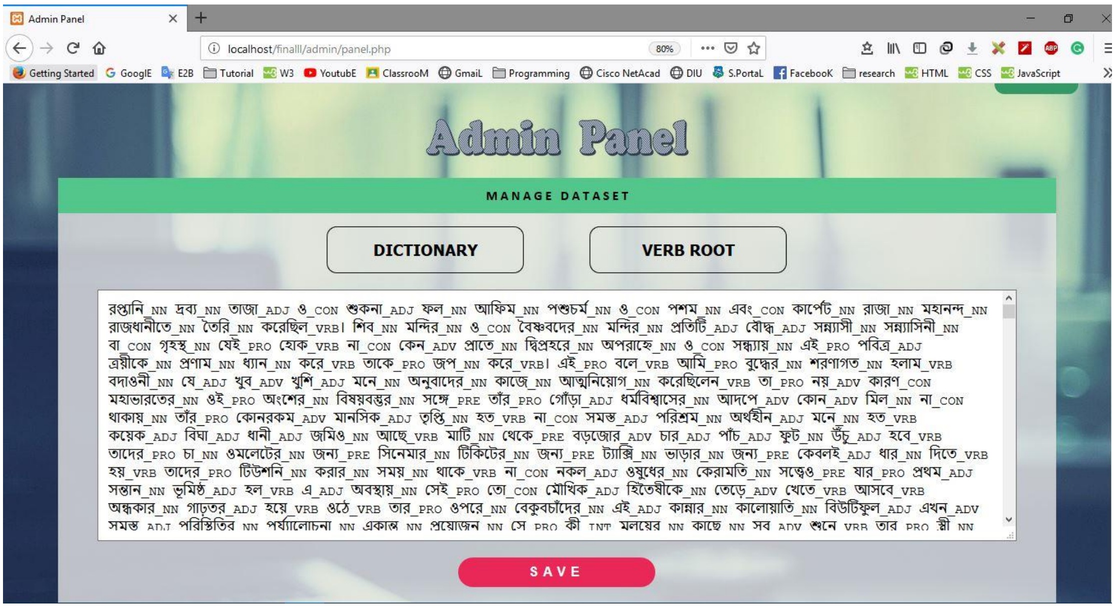


Figure 3.5.2: Login Page for Admin

In figure II, we represent the login page for admin, where user can enter his email and password for login purpose. If input email or password or both are wrong, then an error message will be shown and login will be failed.



```
রপ্তানি_NN দ্রব্য_NN তাজা_ADJ ও_CON শুকনা_ADJ ফল_NN আফিম_NN পশুচর্ম_NN ও_CON পশম_NN এবং_CON কার্পেট_NN রাজা_NN মহানন্দ_NN
রাজধানীতে_NN তৈরি_NN করেছিল_VRB। শিব_NN মন্দির_NN ও_CON বৈষ্ণবদের_NN মন্দির_NN প্রতিটি_ADJ বৌদ্ধ_ADJ সন্ন্যাসী_NN সন্ন্যাসিনী_NN
বা_CON গৃহস্থ_NN যেই_PRO হোক_VRB না_CON কেন_ADV প্রাতে_NN দ্বিপ্রহরে_NN অপরাহ্নে_NN ও_CON সন্ধ্যায়_NN এই_PRO পবিত্র_ADJ
ত্রয়ীকে_NN প্রণাম_NN ধ্যান_NN করে_VRB তাকে_PRO জপ_NN করে_VRB। এই_PRO বলে_VRB আমি_PRO বুদ্ধের_NN শরণাগত_NN হলাম_VRB
বদাওনী_NN যে_ADJ খুব_ADV খুশি_ADJ মনে_NN অনুবাদের_NN কাজে_NN আত্মনিয়োগ_NN করেছিলেন_VRB তা_PRO নয়_ADV কারণ_CON
মহাভারতের_NN ওই_PRO অংশের_NN বিষয়বস্তুর_NN সঙ্গে_PRE তাঁর_PRO গোঁড়া_ADJ ধর্মবিশ্বাসের_NN আদপে_ADV কোন_ADV মিল_NN না_CON
থাকায়_NN তাঁর_PRO কোনরকম_ADV মানসিক_ADJ তৃপ্তি_NN হত_VRB না_CON সমস্ত_ADJ পরিশ্রম_NN অর্থহীন_ADJ মনে_NN হত_VRB
কয়েক_ADJ বিঘা_ADJ ধানী_ADJ জমিও_NN আছে_VRB মাটি_NN থেকে_PRE বড়জোর_ADV চার_ADJ পাঁচ_ADJ ফুট_NN উঁচু_ADJ হবে_VRB
তাদের_PRO চা_NN ওমলেটের_NN জন্য_PRE সিনেমার_NN টিকিটের_NN জন্য_PRE ট্যাক্সি_NN ভাড়ার_NN জন্য_PRE কেবলই_ADJ ধার_NN দিতে_VRB
হয়_VRB তাদের_PRO টিউশনি_NN করার_NN সময়_NN থাকে_VRB না_CON নকল_ADJ ওষুধের_NN কেরামতি_NN সত্ত্বেও_PRE যার_PRO প্রথম_ADJ
সন্তান_NN ভূমিষ্ঠ_ADJ হল_VRB এ_ADJ অবস্থায়_NN সেই_PRO তো_CON মৌখিক_ADJ হিতৈষীকে_NN তেড়ে_ADV খেতে_VRB আসবে_VRB
অন্ধকার_NN গাঢ়তর_ADJ হয়ে_VRB ওঠে_VRB তার_PRO ওপরে_NN বেকুবচাঁদের_NN এই_ADJ কান্নার_NN কালোয়াতি_NN বিউটিফুল_ADJ এখন_ADV
সমস্ত_ADJ পরিস্থিতির_NN পর্য্যালোচনা_NN একান্ত_NN প্রয়োজন_NN সে_PRO কী_INT মলয়ের_NN কাছে_NN সব_ADV শুনে_VRB তার_PRO স্ত্রী_NN
```

Figure 3.5.3: Admin Panel for Manage System Resource Dataset

In figure III, we represent the screenshot of our admin panel, where admin can see and edit the resource datasets (Dictionary dataset and verb root dataset) of the system. There are two options for selecting the dataset. Admin can click on the option, which one admin want to see or edit. Admin can increase the dataset content by adding content at the end of the previous content of the specific dataset.

We must mention that this login and manage dataset options only for admin. User can only interact with the home page. To keep resource privacy, only admin can handle resource content.

Table VII: Admin Database Entity

Entity Name	Type
email	Varchar (100)
password	Varchar (50)

We have only one database table for admin login. In this table, we have two entity, one for admin email and one for admin password, and both are varchar type.

CHAPTER 4
EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental results

The efficiency of a system can be measured from its accuracy level. Our proposed algorithm is applied to the testing dataset, which is collected from different popular online newspaper and blogs. There are almost 12,000 words available in the testing dataset. Accuracy is measured from the ratio of the number of correctly tagged word and the total number of words. Our system can detect 11,304 words with correct tags. It means, our system obtains 94.2% accuracy, which is not a bad figure. Our system detects Verb, Noun, Adjective, and Adverb more efficiently. We experiment on these words in three different contexts. The result are shown in Table VII.

Table VIII: Experiment Result

Total word	Experiment type	Correctly tagged word	Accuracy %
12000	Without dictionary and verb-root dataset	7240	60.3%
	Without dictionary and with verb-root dataset	9831	81.3%
	Dictionary + Verb-root + Rules	11304	94.2%

4.2 Descriptive analysis

By analyzing the result, we identify some constraints. Those are mentioned below-

Firstly, some words wrongly tagged as a Verb. For example – “তাকে”, which is a pronoun. But if we divide the word, we get “তাক্”, which is a Verb root and “ে”, which is a ক্রিয়া-বিভক্তি. According to grammatical rule, it is correct to tag as a verb. But we know, this is Pronoun.

Secondly, there are some foreign words, like “ড্র”, “সিরিজজয়ী”, which counted as a Noun. There are no specific rules can be applied to them.

Thirdly, some English words conflict with some Bangla words. For example – “কার” can be used as to mean Car or to mean Whose. As a result, the word is tagged with the wrong tag.

Fourthly, some non-suffix words, which are not in the dictionary, are automatically counted as a Noun. For example – an Adjective word “চেনা”, which is tagged as a Noun.

Fifthly, some words have different parts-of-speech according to the meaning in the sentence. In this case, maximum occurred tag is applied to them. For example – “ও” is used as Pronoun and Conjunction. Suppose, “ও” is a pronoun in a sentence. But the frequency of Conjunction is high in the dictionary, so “ও” tagged as Conjunction.

Sixthly, word that ends with “ে” and root word is a Noun/Adjective, which is not an Adverb, but still tagged as Adverb. For Example – a word “ফুটবলে”, that ends with “ে”, is tagged as Adverb. But It is a Noun.

Finally, Bangla is a very mysterious language, where each word has multiple forms. So, identify each word with an appropriate tag is not possible. Moreover, Bangla language is influenced by various foreign language. So, many grammatical rules are not applicable to them.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Summary and Conclusion

We successfully develop our proposed Bangla POS tagger system with a satisfying accuracy level (94.2%) using Bangla grammar suffix rules. This POS tagger is performs slightly better than other available POS tagger (Shown in Table VIII). Here dictionary dataset and Verb root dataset helps to increase efficiency level. It can be more efficient. But we don't apply any rule for Pronoun, Conjunction, and Interjection. Because these are the non-suffix words and cannot be recognized without knowing their roles in the sentence. Moreover, Bangla has a huge range of vocabulary, which is influenced by various languages. It is not possible to give an accurate tag to every word. But we can increase the accuracy, if it is possible to find the role of a word in a sentence.

Table IX: Accuracy of Different POS Tagger

POS Tagger	Ekbal [6]	Mukherjee [9]	Parikh [12]	Hoque [14]	Our System
Accuracy %	86.8%	93.1%	92.1%	93.7%	94.2%

5.2 Future work

In this system, we identify tag according to Bangla grammar, especially using suffix analysis. This system mainly focuses on word-level tag accuracy. In the future, we will go for sentence-level accuracy. We will analyze the sentence pattern to identify the role of the word. Here we assign only eight POS tags, but we will try to cover all sub-categories of each base POS tag.

References

[1] S. Mall and U. C. Jaiswal, “Evaluation for POS tagger, chunk and resolving issues in word sense disambiguate in machine translation for Hindi to English languages”, in *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016.

[2] Y. A. Rahman, M. A. Sohan, K. I. Zinnah and M. M. Hoque, “A Framework for Building a Natural Language Interface for Bangla”, in *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2017, pp. 935-940.

[3] C. D. Manning, “Part-of-speech Tagging from 97% to 100%: Is It Time for Some Linguistics?” in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2011, pp. 171-189.

[4] A. R. Pal, N. S. Dash and D. Saha, “An Innovative Lemmatization Technique for Bangla Nouns by using Longest Suffix Stripping Methodology in Decreasing Order” in *International Conference on Computing and Network Communications (CoCoNet)*, 2015, pp. 675-678.

[5] M. F. Kabir, K. Abdullah-Al-Mamun and M. N. Huda, “Deep Learning based Parts of Speech Tagger for Bengali”, in *International Conference on Informatics, Electronics and Vision (ICIEV)*, 2016, pp. 26-29.

[6] A. Ekbal and S. Bandyopadhyay, “Part of Speech Tagging in Bengali using Support Vector Machine”, in *International Conference on Information Technology (ICIT)*,2008, pp. 106-111.

[7] A. Ekbal, S. Mandal and S. Bandyopadhyay, “POS Tagging Using HMM and Rule-based Chunking”, in *Proceedings of the IJCAI, Workshop On SPSAL*, 2007, pp. 25-28.

[8] A. Ekbal, R. Haque and S. Bandyopadhyay, “Maximum Entropy Based Bengali Part of Speech Tagging”, in *Advances in Natural Language Processing and Applications, Research in Computing Science*, vol. 33, 2008, pp. 67-78.

[9] S. Mukherjee and S. K. D. Mandal, “Bengali Parts-of-Speech Tagging using Global Linear Model”, in *Annual IEEE India Conference (INDICON)*, 2013.

[10] S. Dandapat, S. Sarkar and A. Basu, “Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario”, in *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 221–224.

[11] H. Ali, “An Unsupervised Parts-of-Speech Tagger for the Bangla language”, in *Department of Computer Science, University of British Columbia*, 2010.

[12] A. Parikh, “Part-Of-Speech Tagging using Neural network”, in *7th International Conference on Natural Language Processing (ICON)*, 2009.

[13] S. Ismail, M. S. Rahman and Md. A. A. Mumin, “Developing an Automated Bangla Parts Of Speech Tagged Dictionary”, in *16th International Conference on Computer and Information Technology (ICCIT),* 2014, pp. 355-359.

[14] Md. N. Hoque and Md. H. Seddiqui, “Bangla Parts-of-Speech Tagging using Bangla Stemmer and Rule based Analyzer”, in *18th International Conference on Computer and Information Technology (ICCIT),* 2015, pp. 440-444.

[15] Dr. H. Mamud, “ভাষা-শিক্ষা বাংলা ভাষার ব্যাকরণ ও রচনারীতি”, *The Atlas Publishing House,* 2009.

[16] M. Chowdhury and M. H. Chowdhury, “বাংলা ব্যাকরণ ও নির্মিতি”, জাতীয় শিক্ষাক্রম ও পাঠ্যপুস্তক বোর্ড, 2013.

[17] A. T. Dev, “STUDENTS’ FAVOURITE DICTIONARY (BENG. To ENG.)”, Faizuddin, 2004.