Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2022, 10, 43-57
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Construction of an Automatic Bengali Text Summarizer Using Machine Learning Approaches

Busrat Jahan[1], Mahfuja Khatun[2], Zinat Ara Zabu[1], Afranul Hoque[1], Sayed Uddin Rayhan[1]

[1]Department of Computer Science & Engineering, Feni University, Feni, Bangladesh
[2]Department of Computer Science & Engineering, United International University, Dhaka, Bangladesh
Email: hossenbipasa980@gmail.com, mahfuja.duet22@gmail.com, zinatarazabu1997@gmail.com, ahrafi4554@gmail.com, sayedrayhan10@gmail.com

How to cite this paper: Jahan, B., Khatun, M., Zabu, Z.A., Hoque, A. and Rayhan, S.U. (2022) Construction of an Automatic Bengali Text Summarizer Using Machine Learning Approaches. *Journal of Data Analysis and Information Processing*, **10**, 43-57.
https://doi.org/10.4236/jdaip.2022.101003

Received: November 5, 2021
Accepted: February 6, 2022
Published: February 9, 2022

Abstract

In our study, we chose python as the programming platform for finding an Automatic Bengali Document Summarizer. English has sufficient tools to process and receive summarized records. However, there is no specifically applicable to Bengali since Bengali has a lot of ambiguity, it differs from English in terms of grammar. Afterward, this language holds an important place because this language is spoken by 26 core people all over the world. As a result, it has taken a new method to summarize Bengali documents. The proposed system has been designed by using the following stages: pre-processing the sample doc/input doc, word tagging, pronoun replacement, sentence ranking, as well as summary. Pronoun replacement has been used to reduce the incidence of swinging pronouns in the performance review. We ranked sentences based on sentence frequency, numerical figures, and pronoun replacement. Checking the similarity between two sentences in order to exclude one since it has less duplication. Hereby, we've taken 3000 data as input from newspaper and book documents and learned the words to be appropriate with syntax. In addition, to evaluate the performance of the designed summarizer, the design system looked at the different documents. According to the assessment method, the recall, precision, and F-score were 0.70, 0.82 and 0.74, respectively, representing 70%, 82% and 74% recall, precision, and F-score. It has been found that the proper pronoun replacement was 72%.

Keywords

Natural Language Processing, Formatting, Bangla Text Summarizer, Bengali Language Processing, Word Tagging, Pronoun Replacement, Sentence Ranking

1. Introduction

This template, created in MS Word 2007, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: 1) to ease of use when formatting individual papers; 2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products; and 3) conformity of style throughout a journal paper. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow. It is a scientific fact that our universe is expanding over time with the data size in our internet or virtual world expanding more and more enormously. More and more data and documents are being added to this virtual universe every day. However, time is very short and people do not have enough time to verify and read all the documents. The approximate size of the websites that contain e-content, as of July 8, 2020, there were around 5.57 billion pages [1] and it is growing rapidly in each and every second [2]. Thus, science needs to invent a new technology that can reduce the size of documents and make it easier for the world so that they can easily keep a summary of their desired documents. Automatic text summarization is the process where we can meet this demand [2]. It offers to understand a big volume of information in very little time. The automatic English text summarization technique by using the term frequency was first suggested by Luhan about five decades ago [3]. After the upgradation of online volume, Edmundson [4] proposed significant development in summarizing the English text by considering text titles, cue-words, and the position of sentences. Yet, this text summarization trend is not just continuing not only for English but also for Bengali; and nowadays, Bangla magazine of an online portal and electronic Bangla text of Bangla News Documents is expanding rapidly. Only a few studies have been conducted to improve the vast amount of text in Bangla [5] [6] [7]. We know that in the Indo-European language [8], Bangla is the 4th largest language and in the world's number of native speakers' terms, the position is the 6th. It is also the world's 7th-largest oral language (out of 3500 languages) [7] [8]. In the Bangladeshi nation, their mother tongue is Bangla and Bangla is the second most widely spoken language in various Indian states [9]. As stated by economic surveys 2015, the majority of Bangladesh's educated population adapted to the Bengali language is only 62.2% [9].

Moreover, there are few scholars working on Bangla language processing, which is why the possibility of knowledge sharing is limited despite these difficulties, a procedure for Bangla text summarization has been presented here which also focuses on the difficulty of dangling pronouns in summary. As the result of

text summarization, the existence of any dangling pronoun makes disjoint information. Thus, systems are designed to reduce the problem of massive text that can send the erroneous message if there is any dangling pronoun. Users will often be misled by incorrect information without receiving any direction. In these circumstances, we have suggested a method with some major contributions as follows:

- Reducing dangling pronouns in total the number of the summary, replace pronouns with the corresponding nouns;
- Introducing sentence frequency for sentence ranking and eliminating redundancy;
- Identifying numerical figures from the variety of forms (presented in words and digits) to assess the importance of sentences;
- Recognize pronouns and the difference between subject and object;
- Finding the role of every word in Bangla text.

2. Methodology

The proposed Bangla news document summarization approach consists of four main modules: pre-processing; word tagging; replacing pronouns; Sentence ranking with summarization generation respectively. The proposed method's whole procedure is depicted in the following Figure 1.

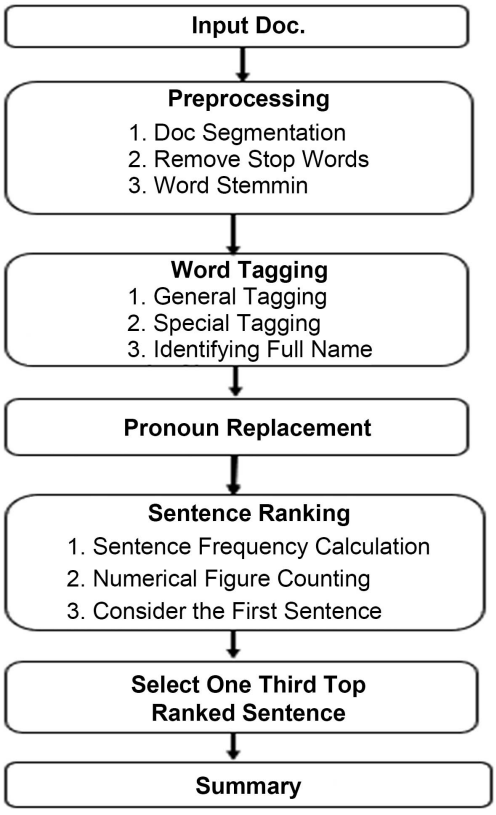

Figure 1. Whole procedure of the proposed automatic Bengali document summarization system.

2.1. Pre-Processing

Pre-processing of the input document is the starting point of this proposed method. Different types of user inputs are needed from one method to another as per the structural design and functional procedure. The preprocessing of the input document in this research work includes: 1) segmenting the input document; 2) removing stop words; and 3) word stemming.

2.2. Word Tagging

General Tagging, we tested 200 Bangla newspaper and discovered that 70% of the terms could be categorized using these prepared lists of words. Though we utilize word stemming to determine the root form of a word, the verb cannot be stemmed when it is not in its active form. In fact, it is very difficult to identifying verb because there are many suffixes in Bangla [10]. The English words "do" can be "doing," "did," or "does" according on the tense and person, but the word "do" can also be "doing," "did," or "does" in Bangla. Take into account the present continuous tense. For example, the three major forms of the word "কর" (kor-do) can only be determined by the first, second, and third person. Also, it can be "করছি" (doing) for first person, "করছ" (doing) for second person and "করছেন" (doing) for third person respectively. Thus this word "কর" (do) may have the next forms: "করে" (do), "করেন" (do), "করিস" (do), "করি" (do), "করছে" (doing), "করছেন" (doing), "করছ" (doing), "করছিস" (doing), "করছি" (doing), "করেছে" (did), "করেছেন" (did), "করেছ" (did), "করেছিস" (did), "করেছি" (did), "করুক" (do), "করুন" (do) and so on. However, verb identification plays a vital role for language processing because this is the main root of a sentence. A list of suffixes are considered as for the final checking in following: "ইতেছিস" (itechhis), "তেছিস" (techhis), "ইতিস" (it is), "ইলে" (ile), "ইবি" (ibi), etc. The percentage of words tagged has increased from 68.12 percent (before using Edmundson's [4] list of suffixes) to 70% (after using Edmundson's list of suffixes).

Special Tagging: Special tagging was created after general tagging to recognize words such as acronym, elementary form, numerical figure, repeating words, name of occupation, organization, and places. The research was conducted out on a total of 32,143 words from 200 different test texts. After considering suffixes for verbs in general tagging rose from 68.12% to 70.00% (after considering suffixes for verbs). In **Table 1** shows several exploratory findings of special tagging are given here: Here, excellent relates to 100%, best relates to 96% - 98%, better relates to 85% - 90%. **Table 2** shows that after special tagging of words has been improved from 70.00% (before considering special tagging) to 76.98% (after considering special tagging) in our experimental result.

Identifying the full names of human for the detection of subjects and objects: in the previous steps, the whole process of tagging somehow depends on a list of words. Point to be mentioned that the existing procedure of named entity recognition [11] has not been utilized in our proposed method. Because it is founded on a predefined list of words only and the impact of surrounding words

Table 1. Exploratory different types of word results of special tagging.

Kinds of term	Result status
English acronym	Best
Repeated words	Excellent
Bengali name	Excellent
Digit	Excellent
Occupation	Best
Places name	Best
Organization name for both cases	Better

Table 2. Experimental results of word tagging at different phases.

Types of word tagging	Total words for tagging	The percentage of word tagging
General tagging	21,896	68.12%
Considering suffixes for the verb	22,500	70.00%
Special tagging	26,098	76.98%

is ignored. According to our analysis, we find that the result of full names of humans is around 95% times, and then parts of the names may be used anywhere in the news documents. There is very tough to get the complete term or name by using a part of a term or name. As a one word may be presented several actions such as, "চাঁন" (Chan) can specify on behalf of "moon" on the other hand "চাঁন মিয়া" (Chan Miah) shows a person's name as a recognized worthy last name "মিয়া" (Miah).

2.3. Pronoun Replacement

While human name recognition is an important task, some other processes are also essential for pronoun replacement. For instance, the subject and object need to be identified from the named entity. In this step, to recognize the subject and object of each sentence, some similar nouns are separated and some rules are applied to replace the pronoun like this: For replacement, eight forms of single pronouns are taken into account from the input document are considered: "তিনি" (tini-she/he), "তাকে" (take-her/him), "তাহাকে" (her/him), "সে" (she/he), "তিনি" (tini-she/he), "উনি" (uni-she/he), "তার" (tar-her/his) and "তাহার" (her/his). Apart from these eight forms, other forms of pronouns have been considered and left as future work, because it will make the process very complex to handle plural forms of pronouns. In our experiment, the last two instantaneous entity sentences are supposed to be a noun related to the pronoun. Our experiments have found that the analogous noun is obtainable 88.63% of the time between the two immediately preceding sentences. When the named entity is not found in the earlier lines then we have applied some rules to search the next earlier lines of the named entity. An illustration of the following textual content is shown to demonstrate the performance of our suggested technique in which the example features and names are imagined.

Sample text: সংবাদ সংস্থার প্রধান আব্দুর রাহিম সাহেব বলেন যে, সকল মানুষের প্রয়োজনীয় খবরের কাগজে ছাপানো উচিৎ।তিনি জানান ভবিষ্যতে পত্রিকা হবে গণমানুষের।

Text after pronoun replacement: সংবাদ সংস্থার প্রধান আব্দুর রাহিম সাহেব বলেন যে, সকল মানুষের প্রয়োজনীয় খবরের কাগজে ছাপানো উচিৎ। আব্দুর রাহিম সাহেব জানান ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য।

Here, making use of our method in the text, the pronoun "তিনি" (tini-he) has been exchanged effectively for the initial multiple times by relating noun "আব্দুর রাহিম" (Abdur Rahim). From the evaluated 200 documents we have counted the pronoun manually and crosschecked it with our program. The results of replacement of pronouns and number of pronouns have been given in the following Table 3.

According to the result in Table 3, our procedure can replace 215 pronouns correctly from 301 Pronouns. So, the accuracy of pronoun replacement is 71.42%.

2.4. Sentence Ranking and Summary Generation

For sentence ranking, values of some attributes are calculated for all the sentences and then sum-up all the attributes' values to compute the final score of each sentence. Top scored sentences are assumed as top-ranked sentences and vice versa. The following attributes are considered in this method: 1) in both words and digits are presented the numerical figure; 2) title words; and 3) the first sentence.

Counting of numerical figures expressed in digits and terms: the first attribute is to count numerical figures for each sentence (SN). The value of SN for each sentence is set to 0 (zero) at first and for the existence of each numerical figure, it will be incremented by 1 (one). A numerical figure (in digits) [12] has been counted and shown to be important for the numerical figure containing a phrase [13]. But it is possible to present the numerical figure in terms that cannot be readily defined as digits. Even, numerical figures can have various suffixes in Bangla text. Therefore, a method is used to recognize the numerical figure of Bangla from both words and digits specified in the special tagging section. All the sentences are segmented into words [w1s1, w2s1, w1s2, w2s2, wnsn] in the pre-processing step and count in digits and terms, the numerical figure based on the following Equations (1)-(3):

$$\forall i \in \{1,\cdots,n\} \quad \text{Ndigits}(i) = \text{Regexp}(\text{S}(i),[0\text{-}9]) \tag{1}$$

$$\forall i \in \{1,\cdots,n\} \quad \text{Nwords}(i) = \text{Regexp}(\text{S}(i),[\text{Format Of Num In Words}]) \tag{2}$$

$$\forall i \in \{1,\cdots,n\} \quad \text{SNC}(i) = \text{Ndigits}(i) + \text{Nwords}(i) \tag{3}$$

Table 3. Experimental results of replacement of pronoun.

Overall pronoun	Untagged	Improperly replaced	Properly replaced
301	71	15	215

where n is the number of sentences; Ndigits and Nwords are the number of numerical figures presented respectively in digits and words; Regexp function returns the number of matches between the corresponding sentence and the given pattern as the second argument of this function. The pattern for matching digits is 0 to 9 and for words is format of num in words (explained in special tagging part). Finally, both Ndigits and Nwords are summed up for each sentence individually to get SN which is the score of numerical figure phrases.

Computation of Score for Title Words (ST): title terms have been considered for sentence scoring in many established methods. In most instances, we have also found from the study of 3000 news documents that title words express the topic of the news article. Every sentence's score for the title word is set to 0 (zero) at first and incremented by 1 for the existence of each title word in the sentence. For computing the title words score of any sentence ST, the title has been segmented to an array of words TW [tw1, tw2, twn] and then proceeds as the following Equation (4):

$$\forall i \in \{1, \cdots, n\} \quad \mathrm{ST}(i) = \mathrm{match}\left(\mathrm{SW}(i), \mathrm{TW}\right) \tag{4}$$

Here, n is the sentence number of the input text and also Sw(i) the array of words for ith sentence, TW is the array of title words, and match function returns the number of words matched with Sw(i) and TW.

Special Consideration of the First Sentence: In some existing methods [12] [13] [14], the sentence score depends on the position where the positional score for the first sentence is the highest and the last sentence is the lowest. From the first sentence, this score decreases steadily. But, most of the time, the first sentence is much more relevant than any other sentence, particularly for Bangla news documents, as per our experiment, which is explained in the lower part of this subsection. In the experiment with our training data set, the first sentence was found to exist 78% of the time in the summary. Therefore, if the first sentence is always held in synthesis, 22 percent (100-78) times would be wrongly chosen. But, after scrutinizing one step forward, it has been discovered that if the first sentence includes any title phrase, 88% of the error rate is 12% when the error rate is summarized (100-88). So, if it includes any title phrase, it is suggested here to pick the first sentence in summary. Point to be noted that this type of special care for the first sentence has been proposed here for the Bangla news documents only and it may not be suitable for others. After measuring all the attributes value, the ultimate rank of each sentence is computed using the following Equation. (5), here the rank of kth sentence is Sk:

$$\begin{aligned} Sk &= w1 \times \mathrm{STF}(k) + w2 \times \mathrm{SSF}(k) + w3 \times \mathrm{SNC}(k) + w4 \times \mathrm{ST}(k), \\ &\text{if } k > 1, \max(Sk) + 1, \text{ if } k = 1 \text{ and } S1 \text{ contains any title word} \end{aligned} \tag{5}$$

where: $0 \le w1, w2, w3, w4 \le 1$; $k = n, n-1, n-2, \cdots, 1$ and n are the number of sentences. The rank of the first sentence will be set as the highest value +1 if it contains any title word so that it will be selected always. After generating all the summaries from the training documents, the average F-score for each value of

the coefficient is calculated by comparing the system-generated summary and the corresponding ideal summary.

2.5. Generate Summary

One-third of the top-rated sentences are extracted as summary sentences after sentence ranking, as in the following Equation (6):

$$\forall i \in \{1, n/3\} \quad \text{SumSen} = \text{SumSen} \bigcup \text{ExtTopScored}(S) \tag{6}$$

where: n is the number of sentences; ExtTopScored function extract top scored sentences from sentence set S; SumSen is the set of summary sentences. The number of summary sentences is kept as approximately one third of the total sentences according to the ratio of input document to summary based on [15] if it is not specified by user.

3. Experimental Result and Discussions

3.1. Data Sets

For training and evaluation of the proposed method, 2200 collected Bangla news documents (each document contains 18 to 25 lines of Unicode text) collected one of the famous Bangladeshi newspapers the daily Prothom-Alo (February 2020). Different types of news are covered a wide range of subjects such as politics, sports, crime, economy, environment, etc. by such kinds of newspapers. We analyze 2000 documents to understand the structure of sentences in news documents and identify the rules for replacing a pronoun with a corresponding noun. For other 200 news documents, three human judges have generated a summary of every document. Summaries created by humans are regarded as reference/model summaries. These 200 documents with appropriate model summaries are viewed as a collection of performance evaluations. For the evaluation of the proposed text summarization method, the output evaluation collection was completed as well as the efficiency measurement of the process of replacing pronouns.

3.2. Evaluation

It is very much needed to check how much efficient is our Bangla Text Summarizer. But it is very much tough and difficult job to do since it is still possible to achieve the sophisticated way. It is very tough to gain a system to asses out developed Bangla Text summarizer, in this circumstance, various types of techniques have been used to assess the summary efficiency, which depends on the: 1) efficient content selection; and 2) presentation quality which may be measured on the basis of grammatical correctness and coherence.

3.3. Experiments and Results

The output summary, sentence scoring figure (shown in **Figure 2**) and mean deviance figure (shown in **Figure 3**) of some sample input documents within our total documents is shown as follows:

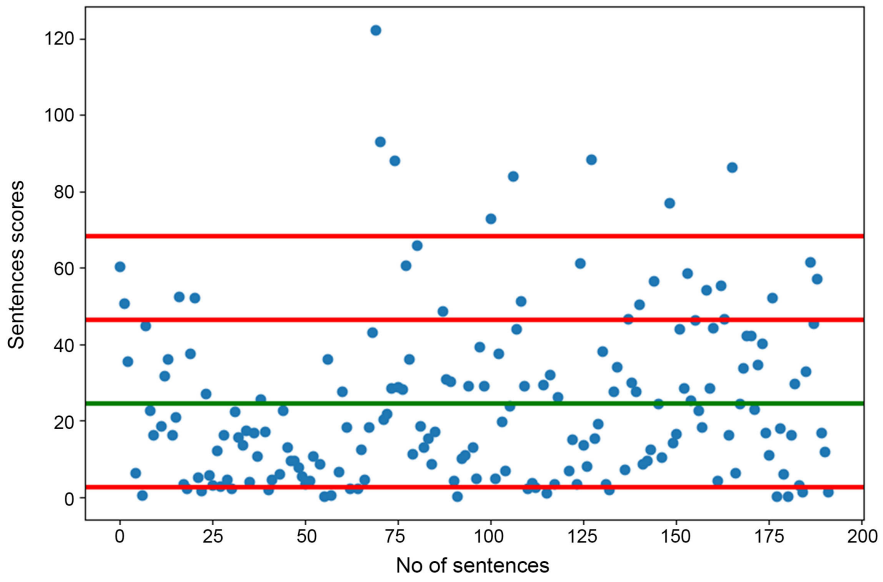

Figure 2. Sentence scoring of sample input.

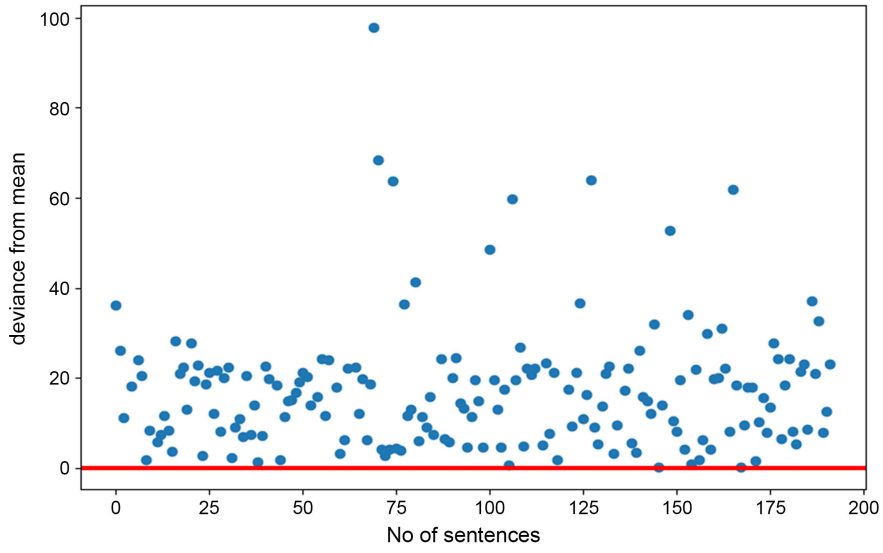

Figure 3. Mean deviance of sample input.

- Sample Input

Title: বুয়েট বন্ধ, হল ত্যাগের নির্দেশ

Text: বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (বুয়েট) আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে। আজ বিকেল পাঁচটার মধ্যে আবাসিক হলে অবস্থানরত সব ছাত্র-ছাত্রীকে হল ছাড়ার নির্দেশ দেওয়া হয়েছে। আজ আড়াইটার দিকে এ আদেশ সংবলিত বিজ্ঞপ্তি বিভিন্ন হলের নোটিশ বোর্ডে সেঁটে দেওয়া হয়।বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত ওই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে।

এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ও কর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষা কার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো। পরীক্ষা পেছানোর দাবিতে 'বিশৃঙ্খলার' প্রেক্ষাপটে বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়-বুয়েট অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করেছে কর্তৃপক্ষ।সেই সঙ্গে বৃহস্পতিবার বিকাল ৫টার মধ্যে ছাত্র-ছাত্রীদের হল ছাড়ার নির্দেশ দেওয়া হয়েছে বলে বুয়েটের ছাত্রকল্যাণ পরিচালক অধ্যাপক দেলোয়ার হোসেন জানান।

- Human Expert' Summary

Title: বুয়েট বন্ধ, হল ত্যাগের নির্দেশ

Text: বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (বুয়েট) আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে। বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত এ আদেশ সংবলিত বিজ্ঞপ্তি বিভিন্ন হলের নোটিশ বোর্ডে সেঁটে দেওয়া হয়।বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে। এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ও কর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষা কার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো।সেই সঙ্গে বৃহস্পতিবার বিকাল ৫টার মধ্যে ছাত্র-ছাত্রীদের হল ছাড়ার নির্দেশ দেওয়া হয়েছে বলে বুয়েটের ছাত্রকল্যাণ পরিচালক অধ্যাপক দেলোয়ার হোসেন জানান।

- System Generated Summary

Title: বুয়েট বন্ধ, হল ত্যাগের নির্দেশ

Getting Summary of Sample Input:

বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (বুয়েট) আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে। বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত ওই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে।পরীক্ষা পেছানোর দাবিতে 'বিশৃঙ্খলার' প্রেক্ষাপটে বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়-বুয়েট অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করেছে কর্তৃপক্ষ। এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ও কর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষা কার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো।

Co-selection measures: In co-selection measures, the principal evaluation metrics are:

$$1)\quad \text{Precision}(P) = (A \cap B)/A \tag{7}$$

where, B = Human Generated Summary (ideal summary);

A = System Generated Summary

$$2)\quad \text{Recall}(R) = (A \cap B)/B \tag{8}$$

where, B = Human Generated Summary (ideal summary);

A = System Generated Summary

$$3)\quad \text{F-Score} = \left(2 \times P \times R\right) / \left(P + R\right) \tag{9}$$

where, P = Precision; R = Recall.

The Evaluation result of fist 10 document has given below in Table 4.

Comparison of Precision, Recall, and F-Score with existing work: To realize the nobility of our proposed system we have to compare our result with existing such kind of work. For this, we have chosen three recent works: 1) K. Sarkar; 2) M.I.A. Efat; 3) Porimol Chandro [16] and compare them with our proposed system's result.

Figure 4 illustrates the comparison of the precision score between the prop osed method and other methods to summarization is shown. In the data visualized ion, the proposed methodology's accuracy score is 0.82, which means that 82%, the proposed methods of Kamal Sarkar achieved 0.57 which means 57%, MIA Efat achieved 0.60 which means 60%, Primal Chandra achieved 0.80 which means 80%. That means our system achieved a higher Precision score than the other existing system.

Figure 5 illustrates the contrast of the recall score between the proposed systems with other approaches to summarization. There is a recall score of 0.70 in the data visualization of the proposed methodology, which means that 70%, Kamal Sarkar proposed methods obtained 0.67 means 67%, MIA Efat achieved 0.50 which means 50% and Primal Chandra achieved 0.67 which means 67%. That means our system achieved a higher Recall score than the other existing system.

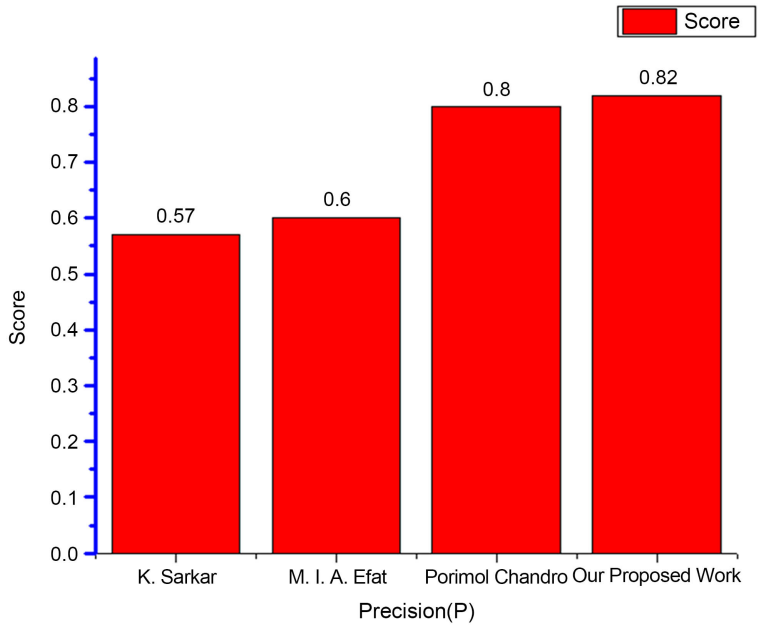

Figure 4. Result of precision score comparison between proposed systems with other summarization approaches.

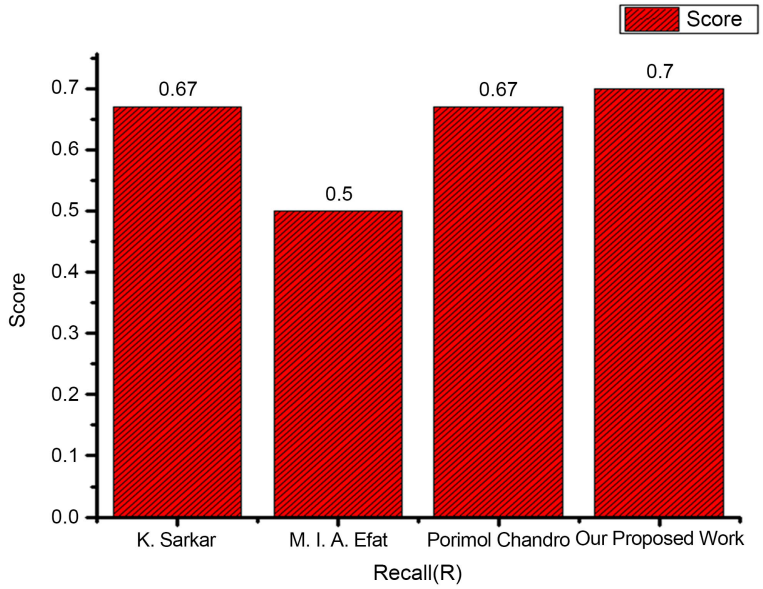

Figure 5. Result of recall score comparison between proposed systems with other summarization approaches.

Table 4. Result of precision, recall and F-Score.

Document No.	Precision(*P*)	Recall(*R*)	F-Score
1	0.84	0.71	0.76
2	0.79	0.72	075
3	0.82	0.69	0.74
4	0.82	0.68	0.74
5	0.79	0.71	074
6	0.82	0.73	0.75
7	0.78	0.72	0.73
8	0.85	0.70	0.75
9	0.85	0.71	0.76
10	0.84	0.71	0.76
Average Score	**0.82**	**0.70**	**0.74**

Figure 6 illustrates the comparison of the F-Score between the proposed method and other methods to summarization is shown. There is an F-score of the proposed methodology in the data visualization of 0.74, which means that 74%, the proposed methods of Kamal Sarkar achieved 0.61 which means 61%, MIA Efat achieved 0.50 which means 50%, and Primal Chandra achieved 0.72 which means 72%. That means our system achieved a higher F-score than the other existing system.

Figure 7 shows the overall comparison of F-Score, Recall & Precision between the proposed method and other methods to summarize.

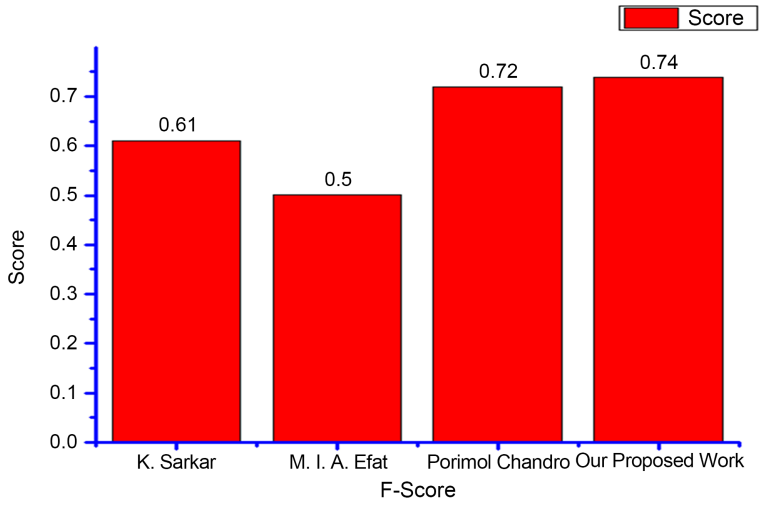

Figure 6. Result of F-Score comparison between proposed systems with other summarization approaches.

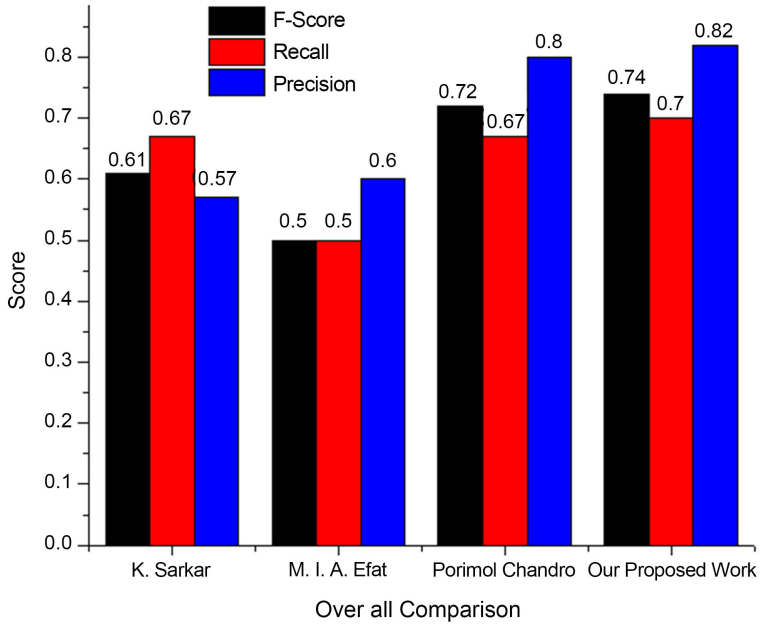

Figure 7. The overall achievement of our and other existing works.

4. Conclusion and Future Work

In our study, an illustration of a new approach in the summary of Bangla news document by introducing an improving version of pronoun replacement and sentence ranking was made. Although there is a lot of research works to summarize English text, which is not directly applicable to Bengali due to the complexity of the Bengali language in terms of sentence structure, grammatical rules, word reflection, etc. In spite of these challenges and barriers, the document was launched in this paper as a groundbreaking way to summarize Bengali news. The proposed method has four steps as follows: 1) preprocessing; 2) word tagging; 3) replacing pronoun by the corresponding noun; and 4) sentence ranking and summary generation.

We have taken over 3000 newspapers and books documents words have been trained according to grammar. And two documents have been checked by the design system to evaluate the efficiency of the designed summarizer. From the evaluation system, it has been found that the recall, precision, F-score are 70%, 82% and 74% respectively, better than the latest existing method. It has been found that the proper pronoun replacement was 72%. There are some limitations of the proposed method also as follows: 1) nature of all the words can't be identified for 100%, 2) though the replacement of pronouns has been introduced, the accuracy of dangling pronouns minimization is 72% and some pronouns are replaced incorrectly. Overall, it can be said that the proposed system obtained potential outcomes as per the results of evaluation, not only for higher ROUGE evaluation scores but also for minimizing dangling pronouns from summary to deliver an unambiguous message. So, it is expected that the proposed system will bring serenity for humans by mitigating the burden of the huge volume of text and lessening the valuable time spent in getting precise information by collecting more images or by image synthesis and augmentation which may help increase the proposed model's accuracy.

Acknowledgements

I would like to express my heartiest gratitude and sincere thanks to my co-authors for their support and encouragement.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] De Kunder, M. (2005) The Size of the World Wide Web.

[2] Ferreira, R. and Luciano, S. (2014) A Multi-Document Summarization System Based on Statistics and Linguistic Treatment. *Journal of Expert Systems with Applications*, **41**, 5780-5787.

[3] Luhn, H.P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, **2**, 159-165. https://doi.org/10.1147/rd.22.0159

[4] Edmundson, H.P. (1969) New Methods in Automatic Extracting. *Journal of the ACM*, **16**, 264-285. https://doi.org/10.1145/321510.321519

[5] Sarkar, K. (2012) Bengali Text Summarization by Sentence Extraction. *Proceedings of International Conference on Business and Information Management* (*ICBIM*-2012), Durgapur, 9-11 January 2012, 233-245.

[6] Sarkar, K. (2012) An Approach to Summarizing Bengali News Documents. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, Chennai, 3-5 August 2012, 857-862. https://doi.org/10.1145/2345396.2345535

[7] Efat, I.A., Ibrahim, M. and Kayesh, H. (2013) Automated Bangla Text Summarization by Sentence Scoring and Ranking. *Proceedings of* 2013 *International Confer-*

ence on Informatics, *Electronics & Vision* (*ICIEV*), Dhaka, 17-18 May 2013, 1-5. https://doi.org/10.1109/ICIEV.2013.6572686

[8] Jahan, B., Emon, I.S., Milu, S.A., Hossain, M.M. and Mahtab S.S. (2021) A Pronoun Replacement-Based Special Tagging System for Bengali Language Processing (BLP). In: Saini, H.S., Sayal, R., Govardhan, A. and Buyya, R., Eds., *Innovations in Computer Science and Engineering*, Springer, Singapore, 761-768. https://doi.org/10.1007/978-981-33-4543-0_80

[9] Farrier, J. (2015) The Second Most Spoken Languages around the World. Olivet Nazarene University, Bourbonnais.

[10] Jahan, B., Mahtab, S.S., Arif, F.H., Emon, I.S., Milu, S.A. and Raju, J. (2021) An Automated Bengali Text Summarization Technique Using Lexicon-Based Approach. In: Saini, H.S., Sayal, R., Govardhan, A. and Buyya, R., Eds., *Innovations in Computer Science and Engineering*, Springer, Singapore, 363-373. https://doi.org/10.1007/978-981-33-4543-0_39

[11] Charniak, E. and McDermott, D. (1985) Introduction to Artificial Intelligence. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

[12] Abuobieda, A., Salim, N., Albaham, A.T., Osman, A.H. and Kumar, Y.J. (2012) Text Summarization Features Selection Method Using Pseudo-Genetic-Based Model. *Proceedings of the* 2012 *International Conference on Information Retrieval Knowledge Management*, Kuala Lumpur, 13-15 March 2012, 193-197. https://doi.org/10.1109/InfRKM.2012.6204980

[13] Sarkar, K. (2014) A Keyphrase-Based Approach to Text Summarization for English and Bengali Documents. *International Journal of Technology Diffusion*, **5**, 28-38.

[14] Baxendale, P.B. (1958) Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development*, **2**, 354-361. https://doi.org/10.1147/rd.24.0354

[15] Radev, D.R., Hovy, E. and McKeown, K. (2002) Introduction to the Special Issue on Summarization. *Computational Linguistics*, **28**, 399-408. https://doi.org/10.1162/089120102762671927

[16] Chandra, P., Arif, F., Rahman, M., Siddik, S., Rahman M.S. and Rahman, A. (2018) Automated Bengali Document Summarization by Collaborating Individual Word &Sentence Scoring. 2018 21*st IEEE International Conference of Computer and Information Technology* (*ICCIT*), Dhaka, 21-23 December 2018, 1-6. https://doi.org/10.1109/ICCITECHN.2018.8631926