Comparative Analysis of Different Text Summarization Techniques Using Enhanced Tokenization

Tanzirul Islam*, Mofazzal Hossain†, MD. Fahim Arefin‡
Dept of Computer Science and Engineering, Green University of Bangladesh, Bangladesh*†‡
Email: tanziruljoy*@gmail.com, mdmofazzalhossainhossain01†@gmail.com, fahim‡@cse.green.edu.bd

2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI) | 978-1-6654-0009-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/STI53101.2021.9732589

***Abstract*—As a huge amount of data is being generated everyday, text summarization is a must-have technique to obtain the required information concisely. Summaries reduce reading time. When it comes to researching documents, summaries make the job easier. The challenge of creating a short and fluent summary while retaining important information content and overall meaning is known as automatic text summarization. As a huge amount of data is being generated everyday, text summarization is a must-have technique to obtain the required information concisely. It is simple to deal with summarization in the other languages like English, Turkish, Arabic. But due to the diverse and complex nature of the Bangla language, not much has been done on the technique of summarizing the Bangla text. Given the importance of text summarization, this paper focused on the creation of an extraction-based summary approach that works on Bangla text documents. Here we apply different kind of model for generating a summary for a single bangla text document. As compared to other outcomes, our experimental results are outstanding and people who read the summary evaluated them. Further development of these methods will undoubtedly deliver fascinating results. This can also contribute significantly in the effort to build smart machines, which form the basis of industry 4.0.**

***Index Terms*—Natural Language Processing, Text Pre-processing, Cosine Similarity based summarization, Extractive summarization, Text Rank based summarization, Word count and heapQ based summarization**

I. Introduction

Natural Language Processing (NLP) is a type of artificial intelligence that enables computers to read, comprehend, and interpret human language. It assists computers in measuring sentiment and determining which portions of human language are significant. However, because of the vast amount of unstructured data, the lack of formal rules, and the lack of real-world context or intent, human language is exceedingly difficult for computers to interpret. Automatic text summarization is the task of producing a concise and fluent summary while preserving the key information content and overall meaning. It is challenging for machines because humans can readily choose the keywords from a given text and summarize them, but computers cannot. As summarization requires linguistic comprehension, reasoning, and the application of common sense information in the same way that humans do, text summarization is quite complex. Summaries facilitate the selection process since they minimize reading time while investigating documents.[1] As a result, there is a strong need to simplify much of this text material into shorter, focused summaries that capture the essential information, both so we can traverse it more efficiently and to determine whether the bigger papers include the information we want. Text can be summarized in two ways: extractive and abstractive. [2]

In extraction-based summarizing, a subset of words representing the most significant points from a piece of text is extracted and merged to form a summary. Consider it similar to a highlighter, which chooses the most important information from a source text. Advanced deep learning algorithms are used in abstraction-based summarization to paraphrase and reduce the original material in the same way that people do.[3] Consider it a pen that generates fresh sentences that may or may not be part of the underlying material. Bangla is one of the world's most widely spoken languages. However, text summarization in Bangla has been limited by various of issues, including the lack of digital standard text databases and inflectional morphology in Bangla. As a consequence, we performed an analysis of several text summarization approaches utilizing enhanced tokenization and obtained some interesting findings. To obtain the desired outcome, we must first go through certain procedures like tokenization, stop words removal, POS tagging, and model construction.

Where The initial step in any NLP pipeline is to tokenize the data. [4] It has a significant impact on the entire pipeline. A tokenizer is a program that converts unstructured data and natural language text into discrete information pieces. Stop word removal is one of the most often utilized preprocessing stages in various NLP applications. The goal is to simply remove words that appear often across all of the documents in the corpus and POS tagging is the process of transforming a sentence into a form – a list of words, a list of tuples (each tuple having a form (word, tag)). The tag in the case of is a part-of-speech tag that indicates whether the word is a noun, adjective, verb, or other.

We deal with all of these phases in our technique. We

implemented most of the steps of text summarization in order to increase efficiency and produce a more fluent and effective summary.

II. Literature review

The last several years a large amount of work have been done with extractive method and many papers have been published on it. And Bengali being the 7^{th} most spoken language all over the world.

Shofi Ullah, Sagar Hossain, K. M. Azharul Hasan [5] worked on cosine simillarity based graph ranking and Relevance based scoring and ranking approach for the summarization of bangla text. And the Method consist of: POS tagging, redundancy removal algorithm & cosine similarity Based graph. Sarkar [6] proposed a summarization technique implementing term frequency and semantic sentence similarity based summarizing approach to summarize a single Bangla text. This method consisted of four major steps: preprocessing, stemming, sentence ranking and summary generation. Sentences are ranked based on two important features: thematic term and position. In paper [7], the reasearchers take a document as input, find the word frequency, determine the sentence similarity, weight the sentences, sort the sentences according to their rank and finally select the sentences with the higher rank for generating the summary.

Talukder et. al. [8] worked with bi-directional RNNs with LSTM in encoding layer and attention model at decoding layer. In this model, they implement an abstractive summarizer and reduce the given text and run time of the model.And the model works as sequence to sequence model to generate summary. Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, Ayoub Bagheri [9] developed a query-oriented text summarization technique which extracted the most informative sentences. Each sentence had 11 of the greatest qualities taken from it. They used the ROUGE criterion to illustrate that using more appropriate features results in better summaries. They use certain appropriate features, with the first set of features identifying informative sentences and the second set of features assisting in the extraction of query relevant sentences. To improve text summarizing, their system reliably picks the most informative sentences and then provides the summary.

Narendra Andhale, L.A. Bewoor's work [4] includes a clear examination of both the extractive and abstractive approaches to text summarization, as well as their respective limitations. The examined literature points to a difficult area where these methodologies could be used to provide interesting, well-compressed, and legible summaries.

Md Mahbubur Rahman, Md. Aktaruzzaman Pramanik, Rifat Sadik, Monikrishna Roy, Partha Chakraborty [2] used a transformer-based bangla document classification system. They used the most recent transformer or attention mechanism-based models to classify Bangla text content. For Bengali text classification, they used the BERT (Bidirectional Encoder Representations from Transformers) and ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) models. Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal,[10] described a text summarizing algorithm that extracts key lines from a single or many Bengali papers. They used the K-means clustering approach to provide the final summary for single or many articles. They compared the results to those of other extractive summarization techniques and measured the run-time complexity, demonstrating that the new technique's performance has improved. Sheikh Abujar, Mahmudul Hasan, M.S.I Shahin Syed Akhter Hossain [3] focused with the extraction method for summarizing Bangla text. Basic extractive summarization was used for the new proposed model in their proposed amendment. A set of rules for analyzing Bangla text was also developed from the heuristics.The material is carefully examined using the Bangla sentence clustering method. Their methodology groups all of the data together and presents the information in a clear and concise manner. Prachi Shah and Nikita P. Desai[1] examined text summarizing techniques for a variety of Indian and foreign languages, including English, European, and others. They also offered a method for summarizing Hindi text using machine learning techniques. This research focused on surveying and analyzing the performance of automated text summarizers for a variety of Indian and international languages. The top ranked sentences are considered in the summary, and their approach assigns ranks based on sentence score.

In this section, we summarized the state of the art works in the field of text summarization in Bangla. We discuss our proposed methodology in the next section.

III. Proposed methodology

Before summarizing a text, we need to complete a few pre-processing steps, we highlight those in Fig. 1.

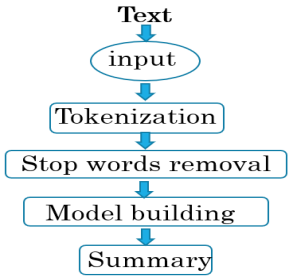

Fig. 1: Work Flow of summary generation

For every approach, we have to complete some preprocessing. Those are tokenization, removing stopwords, and lemmatization.

(a) Tokenization: Tokenization is one of the common task when it comes to working with text data. Tokenization is essentially splitting a phrase, sentence, or an entire text document into smaller units such as individual words or sentence. units are called tokens.

(b) Removing Stopwords: Stop words are the words which have less significance in texts and have very high or very low occurence in the document and they are removed. More than 350 stop words available for bengali texts such as: অতএব, আগামী, অবধি,আর, আরও, ইত্যাদি, এত, এতটাই, অথচ, অথবা

(c) Lemmatization: Lemmatization is a method for combining various inflected forms of words into a single root form with the same meaning. It's similar to stemming in that it results in a stripped-down word with dictionary meaning. There are several words in different formations like বাংলাদশেী, বাংলাদশেরে, বাংলাদশেীয়, বাংলাদশেকেetc. They basically denote the same meaning বাংলাদশে but with the inclusion of terms like ‘s’, ‘es’ it shows the different form. To remove redundancy, the conversion of several formations into an actual word is also needed.

In this paper, we used four different approaches,for generating a summary for a single text document which are as follows:

1) Cosine Similarity based summarization
2) Extractive summarization
3) Text Rank based summarization
4) Word count and heapQ based summarization

A. Cosine similarity-based summarizing

It is mostly used for hierarchical clustering and multidimensional scaling, after which we evaluate sentence similarity and merge the ranking sentences for summary those that are approximately one or one. To begin cleaning, we remove the white space, bangla digits, upper and lower case English letters, and punctuation-marks. After cleaning the text, we split it into sentences and then ranked each phrase on a scale of (0-1) it's would be like

বশি্বকাপে ৮ ম্যাচে ৩৪ ২৮ গড়ে ২৪০ রান করছেলিনে 0.08186748647115961
২টি সঞ্চেুরি ও ৯টি ফফিটি ছলি তাঁর 0.07580509024530348
৪২ ম্যাচে ২৮ ৪৮ গড়ে ৮৮৩ রান করছেলিনে 0.07440322422522412
এ ছাড়া অস্ট্রলেয়িার বপিক্ষওে ৪০ রান করছেলিনে যশপাল 0.07402828995191987
তাতে ছলি চারটি ফফিটি 0.0665731254875899
তাতে ৩৩ ৪৫ গড়ে ১৬০৬ রান করছেলিনে 0.0651854654535155
ছলি দুটি ফফিটি 0.062214838422544744

Here we can see that every sentence are ranked to(0-1) and since we are using the positive quadrant, which means no sentence may have a negative value, before scoring all sentences from the input text. Finally, we compute the summary frequency (how many sentences are produced) from the rated sentences.

B. Extractive Summarization:

There are two methods for extractive summarization: First of all The extractive method will take the same words, phrases, and sentences from the original text. Extractive methods can be considered as important sentence selection in the given text - one is percent summary and the other one is word count summary.

Percent summary: For percent summary we use ratio = 0.5 which means that the ratio between two sentences. * the similarity measure of each sentence with ratio 0.5(50* after that ranked those sentences for summary

Word count summary: for Word count summary we use word-count = 100 which means that the summary will be within 100 words.

- Here we just count those word are repeated many time in a sentence.
- Then select those sentences for summary within 100 words

C. Textrank based summarization

It focuses on the ranking of text sentences and is computed recursively based on information accessible throughout the text. TextRank is a graph-based algorithm, easy to understand and implement. It does the necessary preprocessing before calculating the similarity between texts. Then it ranks the sentences based on their importance before providing the summary. TextRank works in the following steps:

1) Tokenize documents into sentences.
2) Preprocess each sentence in the document.
3) Calculate the Similarity between sentences.
4) Rank the sentences with higher significance.

['মোল্লা মোহাম্মদ হাসান আখুন্দকে প্রধানমন্ত্রী করে ৭ সপ্টেম্বের আফগানস্তিানরে অন্তর্বর্তীকালীন সরকাররে ঘোষণা দয়িছে তালবোন।',
'মন্ত্রসিভার সবাই পুরুষ এবং আগ্নয়োস্ত্রসহ কন্দ্রীয় ব্যাংকরে গভর্নররে ছবি আমরা সামাজকি মাধ্যমে দখেছে।'
, 'তারও আগে ১৫ আগস্ট রাজধানী কাবুলসহ দশেটরি নয়িন্ত্রণ নয়ে তারা।'] This is the first step to tokenize documents into sentences and preprocess each sentence in the text document. The output summary will consist of the most representative sentences and will also be returned as a string, divided by newlines. If the split parameter is set to True, a list of sentences will be returned. The length of the output can be specified using the ratio and word-count parameters:ratio should be a number between 0 and 1 that determines the percentage of the number of sentences of the original text to be chosen for the summary (defaults at 0.2).word-count determines how many words will the output contain. If both parameters are provided, the ratio will be ignored, and finally given the summary for the given text those sentence are higher signficance, and this method is more popular nowadays for both Abstractive and Extractive summarization.

iv.The input text is cleaned first by word count and heapQ-based summarization. Following cleaning, After cleaning the text like (white space,bangla-digits Like(1,2,3,4),english letter (Aa-zZ) both upper case and lower case, punctuation-mark.We remove the stop word then we divide it to sentences,Then scoring all sentence using word count(it means that it count all word of a sentence)it eliminates stopwords, tokenizes the text, and scores all sentences based on word count. Then scoring all sentence we use heapQ for our summary that take the largest value only to make the summary.For example-

মরীচকিা চরকরি আট পর্বরে অরজিনিাল ওয়বে সরিজি 2.75
ওয়বে সরিজিটরি পরচিালক শহিাব শাহীন জানালনে দর্শকরে এই সাড়া তাঁকে আপ্লুত করে 4.5
তনিি বলনে ট্রলোরটি সরিজিরে একটি চুম্বক অংশ 1.75
আমার বশ্বিাস পুরো সরিজি দর্শকরে আরও ভালো লাগবে 3.75
বাংলাদশেরে কনটন্টে হসিবেে এই সরিজি দর্শকরে আলাদাভাবে ভাবাবে 3.25
তবে পুরো কাজ তুলে আনাটা তাঁর জন্য চ্যালঞ্জেং ছলি বলে জানালনে শহিাব শাহীন 4.5
.....

Here we can see that every sentence are tokens with their individual significance number and among them which sentence is provide highest number that sentence will be the part of the summary.

IV. Experimental Result

In this section we discussed all methods input and output as sample text document each and every method have an own text document and we represented all of methods. First, we collected some articles from Prothom Alo newspaper. Then we summarize it, and then we compare each and every method's output to each other, according to what we consider to be each method's summary. We give these outputs to some of our friends and senior brothers to know which output is nearer to a human summary or more accurate as a summary.

In TABLE I, we present a sample result for Cosine Similarity based summarization. At first we take a text document as input and after processing, we get our summary for the given text document.

For Extractive summarization we represented percent summary and word count summary in TABLE II. In percent summary, we use ratio = 0.5 which refers to the ratio between main text and summary length. We can change the ratio as for our given text here we use 0.5 because we use here mid-sized text document. If it became large we reduce the value of it as for our user requirement. It will find the ratio between two sentence from the given text and after that rank those sentences to generate the percent summary. In extractive method, it takes the same words, phrases, and sentences from the original text. For Word count summary, we use word-count = 100 which means that the summary will be within 100 words for the given text document. We can change the value of it as for our user requirement. Extractive summarization is easy to implement

TABLE I: sample result for Cosine Similarity based summarization

Input	সাবকে ভারতীয় ক্রকিটেোর ও ১৯৮৩ বশ্বিকাপজয়ী যশপাল শর্মা আজ মঙ্গলবার মৃত্যুবরণ করছেনে। ৬৬ বছর বয়সী এই বশ্বিকাপজয়ী ক্রকিটেোররে মৃত্যুর কারণ হার্ট অ্যাটাক বলে জানয়িছে ভারতীয় সংবাদমাধ্যম। যশপালস্ত্রী ও তনি সন্তান রখেে গছেনে। সত্তররে দশকরে শষে থকেে আশরি দশকরে মধ্যভাগ পর্যন্ত ভারতীয় দলরে অংশ ছলিনে যশপাল। পাঞ্জাবরে এই মডিল অর্ডাররে অভষিকে হয়ছেলি পাকস্তিানরে বপিক্ষ। ১৯৭৮ সালে শয়িালকোটেরে সে ওয়ানডরে পর আরও ৪১টি ওয়ানডে খলেছেনে যশপাল। ৪২ ম্যাচে ২৮.৪৮ গড়ে ৮৮৩ রান করছেলিনে। তাতে ছলি চারটি ফফিটি। বে যশপাল তাঁর ক্যারয়িাররে সরো মুহূর্ত কাটয়িছেনে ১৯৮৩ বশ্বিকাপে। বশ্বিকাপে ৮ ম্যাচে ৩৪.২৮ গড়ে ২৪০ রান করছেলিনে। ছলি দুটি ফফিটি। প্রথম ম্যাচে যশপালরে ক্যারয়িার-সরো ৮৯ রানরে ইনংিসে ওয়স্টে ইন্ডজিকে চমকে দয়িছেলি ভারত। সমেফিাইনালে ইংল্যান্ডরে বপিক্ষ ৬১ রানরে ম্যাচ জতোনো ইনংিস ছলি তাঁর। এ ছাড়া অস্ট্রলেয়িার বপিক্ষও ৪০ রান করছেলিনে যশপাল। ভারতরে হয়ে ৩৭টি টস্টেও খলেছেনে এই ব্যাটসম্যান। তাতে ৩৩.৪৫ গড়ে ১৬০৬ রান করছেলিনে। ২টি সঞ্চুরি ও ৯টি ফফিটি ছলি তাঁর। প্রথম শ্রণেরি ক্যারয়িারে ১৬০ ম্যাচে ২১ সঞ্চুরতিে ৮ হাজার ৯৩৩ রান যশপালরে।
Output	বশ্বিকাপে ৮ ম্যাচে ৩৪ ২৮ গড়ে ২৪০ রান করছেলিনে। ২টি সঞ্চুরি ও ৯টি ফফিটি ছলি তাঁর। ৪২ ম্যাচে ৮ ৪৮ গড়ে ৮৮৩ রান করছেলিনে। এ ছাড়া অস্ট্রলেয়িার ২বপিক্ষও ৪০ রান করছেলিনে যশপাল। তাতে ছলি চারটি ফফিটি

and understand and it is more familiar for python language. It is more interesting because we get two summary for one method at a time for a single text document. So here is an example of this method and it works on any kind of text document.

In TABLE III and TABLE IV, we try to get a summary with different kind of approaches which is different from other approaches. The output summary will consist of the most representative sentences of the original text for the summary (defaults at 0.2). Here, word-count determines how many words will the output contain to get the summary. Our first step is to tokenize all sentence from the given text and every sentences are ranked with their individual significance number and among them which sentence is provide highest number that sentence will be the part of the summary. And it work for a single text document to generate this summary.

We highlighted our results in Fig. 2. Cosine Similarity based summarization was liked by 55 percent of them, Extractive summarization-70 percent, Text Rank based

TABLE II: sample result for Extractive summarization

Input	মনরে সুখে মায়রে হাতরে হালুয়া খয়েে রীতমিতো ওজন বাড়য়িে ফলেছেনে বলডিডরে সুপারফটি সুপারস্টার অক্ষয় কুমার। মাত্র কদনিইে তার পাঁচ কজেি ওজন বড়েে গছে।ে তবে ওজন বাড়ায় মোটওে বচিলতি নন অক্ষয়। বরং তনিি বজোয় খুশ।ি কারণ, পররে ছবরি প্রয়োজনে এমনতিইে তাঁর ওজন বাড়ানোটা জরুরি ছলি। এই মুহূর্তে আগামী ছবি 'রক্ষাবন্ধন'-এর শুটংিয়ে ব্যস্ত অক্ষয়। আনন্দ এল রাই পরচিালতি এই ছবটিরি শুটংি চলছে মুম্বাইত।ে সনিমোয় তনিি পাঁচ বোনরে একমাত্র ভাই। ভাই-বোনরে সম্পর্করে এক সুন্দর রসায়ন নয়িে এক পারবিারকি ছবি নর্িমাণ করতে চলছেনে আনন্দ। পর্দায় অক্ষয়রে বোনরে চরত্রিে পাঁচ নবাগতা অভনিত্রেীকে দখো যাব।ে জানা গছে,ে এই ছবরি চরত্রিরে প্রয়োজনে অক্ষয়রে ওজন বাড়ানোর প্রয়োজন ছলি। আর তনিি স্বাভাবকি পদ্ধতত িনজিরে ওজন বাড়য়িছেনে। এ জন্য এই বলডিড সুপারস্টার প্রাণভরে খয়েছেনে মায়রে হাতরে হালুয়াসহ আরও নানান বাহারি খাবার। এত দনি যা তনিি মুখে তুলতে পারনেনি। দীর্ঘদনি পর হালুয়া খতেে পরেে দারুণ খুশি অক্ষয়। আর খুব কম সময়রে মধ্যে পাঁচ কজেি ওজন বাড়য়িে ফলেছেনে।
Output	Percent summary:মনরে সুখে মায়রে হাতরে হালুয়া খয়েে রীতমিতো ওজন বাড়য়িে ফলেছেনে বলডিডরে সুপারফটি সুপারস্টার অক্ষয় কুমার। মাত্র কদিনেই তার পাঁচ কজেি ওজন বড়েে গছে।ে কারণ, পররে ছবরি প্রয়োজনে এমনতিইে তাঁর ওজন বাড়ানোটা জরুরি ছলি। সনিমোয় তনিি পাঁচ বোনরে একমাত্র ভাই। পর্দায় অক্ষয়রে বোনরে চরত্রিে পাঁচ নবাগতা অভনিত্রেীকে দখো যাব।েজানা গছে,ে এই ছবরি চরত্রিরে প্রয়োজনে অক্ষয়রে ওজন বাড়ানোর প্রয়োজন ছলি। আর তনিি স্বাভাবকি পদ্ধতত িনজিরে ওজন বাড়য়িছেনে। আর খুব কম সময়রে মধ্যে পাঁচ কজেি ওজন বাড়য়িে ফলেছেনে। Word count summary:মনরে সুখে মায়রে হাতরে হালুয়া খয়েে রীতমিতো ওজন বাড়য়িে ফলেছেনে বলডিডরে সুপারফটি সুপারস্টার অক্ষয় কুমার। মাত্র কদনিইে তার পাঁচ কজেি ওজন বড়েে গছে।ে তবে ওজন বাড়ায় মোটওে বচিলতি নন অক্ষয়। কারণ, পররে ছবরি প্রয়োজনে এমনতিইে তাঁর ওজন বাড়ানোটা জরুরি ছলি। এই মুহূর্তে আগামী ছবি 'রক্ষাবন্ধন'-এর শুটংিয়ে ব্যস্ত অক্ষয়। সনিমোয় তনিি পাঁচ বোনরে একমাত্র ভাই। পর্দায় অক্ষয়রে বোনরে চরত্রিে পাঁচ নবাগতা অভনিত্রেীকে দখো যাব।ে জানা গছে,ে এই ছবরি চরত্রিরে প্রয়োজনে অক্ষয়রে ওজন বাড়ানোর প্রয়োজন ছলি। আর তনিি স্বাভাবকি পদ্ধতত িনজিরে ওজন বাড়য়িছেনে। দীর্ঘদনি পর হালুয়া খতেে পরেে দারুণ খুশি অক্ষয়। আর খুব কম সময়রে মধ্যে পাঁচ কজেি ওজন বাড়য়িে ফলেছেনে।

TABLE III: sample result for Text Rank based summarization

Input	২৮ বছর অপক্ষোর পর পাওয়া পরম আরাধ্য এক শরিোপা। আর্জন্টেনিায় এখন খুশরি জোয়ার। আবগে বাঁধ না মানাই স্বাভাবকি। দলরে মডিফলি্ডার রদ্রগিদো পলও আবগেে লাগাম দতিে পারনেনি। আর্জন্টেনিার হয়ে বাতস্তিুতা-সমিওিনরো সইে ১৯৯৩ সালে যখন মহাদশেসরো হয়ছেলিনে, দ পিলরে তখন জন্মও হয়নি। সমর্থক হসিবেে হোক বা খলেোয়াড় হসিবে, দলরে সাফল্য দখেনেনি কখনো। কোপা আমরেকিা জতোর পর উচ্ছ্বসতি দ পিলরে হয়তো স্বাভাবকি জ্ঞান একটু হলওে লোপ পয়েছেলি। শরিোপা নয়িে উদ্যাপন করতে গয়িে ব্রাজলিকে টনেে আনতে চাইছলিনে। কোপার ফাইনাল দখেতে সাড়ে ছয় হাজাররেও বশেি দর্শককে ঢুকতে দওেয়া হয়ছেলি মারাকানায়, যাঁদরে অধকিাংশই ছলিনে ব্রাজলি-সমর্থক। ভাঙা হৃদয় নয়িে দখেছেলিনে আর্জন্টোইনদরে শরিোপা উৎসব। এই অবস্থায় তাঁদরে কাটা ঘায়ে নুনরে ছটিা দওেয়ার যাবতীয় বন্দোবস্ত করে ফলেছেলিনে দ পিল। উদ্যাপন করতে গয়িে ব্রাজলিবদ্বিষেী এক গান গাইতে শুরু করছেলিনে। কন্িতু মসেি সটো হতে দবেনে কনে। আঙুল উঁচয়িে দ পিলকে সতর্ক করে দয়িছেনে মসেি, যনে ব্রাজলিরে দর্শকদরে সামনে তাঁদরে দশেকে হয়ে করে কোনো গান গাওয়া না হয়। ফুটবলরে জয় -পরাজয় মাঠইে থাকুক, সটোর ওপর ভত্তিি করে দুই দশেরে মধ্যে পারস্পরকি শ্রদ্ধাবোধরে জায়গাটুকু যনে নষ্ট না হয়, সে দকিে সতর্ক ছলিনে মসেি। দ পিলকে সাফ জানয়িে দয়িছেনে, 'আমার চোখরে সামনে এসব গাওয়া যাবে না!' ভডিওিটা সামাজকি যোগাযোগমাধ্যমে ছড়য়িে পড়ছে এর মধ্যইে। অধনিায়ক হসিবেে মসেরি সুনাম চলছে সব জায়গায়। শুধু মসেইি নন, দ পিলকে সইে গান গাইতে মানা করনে আরকে সনিয়ির সতীর্থ সরেহওি আগুয়রোেও। তবে মাঠে দর্শকদরে সামনে না গাইলওে ড্রসেংিরুমে দ পিলকে আটকে রাখা যায়নি। সতীর্থদরে সঙ্গে ওই গান গয়েছেনে, তালে তালে নচেছেনে। সে ভডিওিতে অবশ্য মসেকিে অংশ নতিে দখো যায়নি।
Output	উদ্যাপন করতে গয়িে ব্রাজলিবদ্বিষেী এক গান গাইতে শুরু করছেলিনে। আঙুল উঁচয়িে দ পিলকে সতর্ক করে দয়িছেনে মসেি, যনে ব্রাজলিরে দর্শকদরে সামনে তাঁদরে দশেকে হয়ে করে কোনো গান গাওয়া না হয়। শুধু মসেইি নন, দ পিলকে সইে গান গাইতে মানা করনে আরকে সনিয়ির সতীর্থ সরেহওি আগুয়রোেও। তবে মাঠে দর্শকদরে সামনে না গাইলওে ড্রসেংিরুমে দ পিলকে আটকে রাখা যায়নি।

TABLE IV: sample result for Word count and heapQ based summarization

Input	গত জুনরে শুরুতে ‘চরকি’ ইউটিউব চ্যানলে ‘মরীচকিা’র ট্রলোর প্রকাশরে পরপর ফসেবুকে ছড়য়িে পড়ে। পুরো সরিজি দখোর জন্য ‘মরীচকিা’ চরকরি আট দর্শকরে মধ্যে বাড়তি আগ্রহ তরৈি হয়। পর্ববরে অরজিনিাল ওয়বে সরিজি। ওয়বে সরিজিটরি পরচিালক শহিাব শাহীন জানালনে,দর্শকরে এই সাড়া তাঁকে আপ্লুত করে। তনিি বলনে, ‘ট্রলোরটি সরিজিরে একটি চুম্বক অংশ। আমার বশি্বাস পুরো সরিজি দর্শকরে আরও ভালো লাগবে। বাংলাদশেরে কনটন্টে হসিবেে এই সরিজি দর্শকরে আলাদাভাবে ভাবাবে।’ তবে পুরো কাজ তুলে আনাটা তাঁর জন্য চ্যালঞ্জে ছলি বলে জানালনে শহিাব শাহীন। শুটংিয়ে নমেে নর্িধারতি শডিউিলরে আরও দড়ে গুণ শডিউিল বাড়তি দতিে হয়ছে। সইে যুদ্ধে শল্পিীরাও সহযোগতিার হাত বাড়য়িে দয়িছেনে। মরীচকিার চার অভনিয়শল্পিী মাহি, নশিো, সয়িাম ও জভোন বলনে, ‘শল্পিীদরে সহযোগতিা না পলেে কাজটি হয়তো শষে করতে পারতাম না।
Output	পুরো সরিজি দখোর জন্য দর্শকরে মধ্যে বাড়তি আগ্রহ তরৈি হয়। ওয়বে সরিজিটরি পরচিালক শহিাব শাহীন জানালনে দর্শকরে এই সাড়া তাঁকে আপ্লুত করে। তবে পুরো কাজ তুলে আনাটা তাঁর জন্য চ্যালঞ্জে ছলি বলে জানালনে শহিাব শাহীন। আমার বশি্বাস পুরো সরিজি দর্শকরে আরও ভালো লাগবে

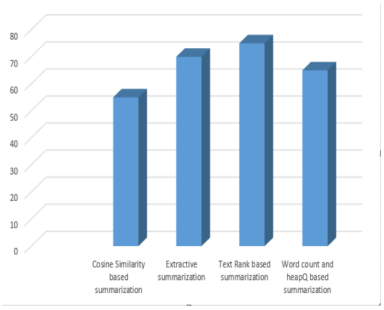

Fig. 2: Result Analysis

summarization-75 percent, and Word count and heapQ based summarization was liked by 65 percent of them.

V. Conclusion

In this paper, we have implemented and compared text summarization using different techniques. To compare, We summarize a single bangla text document with all approaches, namely Cosine Similarity based Summarization, Extractive summarization, Text Rank based summarization, Word count and heapQ based summarization. The first one is based on performing hierarchical clustering and multidimensional scaling after that we check sentence similarity then join the ranked sentence for summary those nearly one or one. In extractive summarization two methods are implemented one is percent summary and the other one word count summary. Textrank based summarization It focused on the ranking of text sentences and computed recursively based on information accessible throughout the text.The input text is cleaned first by word count and heapQ-based summarization, it eliminates stopwords, tokenizes the text, and scores all sentences based on word count. After scoring all the sentences, heapQ was used to get the final summary. In future we improve our work with different kind of approaches and develop an android app for the user.

Acknowledgment

This work was supported in part by the Center for Research, Innovation, and Transformation of Green University of Bangladesh.

References

[1] P. Shah and N. P. Desai, ``A survey of automatic text summarization techniques for indian and foreign languages,'' in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 4598--4601.

[2] N. Andhale and L. Bewoor, ``An overview of text summarization techniques,'' in 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 2016, pp. 1--7.

[3] S. Abujar, M. Hasan, M. Shahin, and S. A. Hossain, ``A heuristic approach of text summarization for bengali documentation,'' in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1--8.

[4] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, ``Query-oriented text summarization using sentence extraction technique,'' in 2018 4th International Conference on Web Research (ICWR), 2018, pp. 128--132.

[5] A. Sarkar and M. S. Hossen, ``Automatic bangla text summarization using term frequency and semantic similarity approach,'' in 2018 21st International Conference of Computer and Information Technology (ICCIT), 2018, pp. 1--6.

[6] S. Ullah, S. Hossain, and K. M. Azharul Hasan, ``Opinion summarization of bangla texts using cosine simillarity based graph ranking and relevance based approach,'' in 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019, pp. 1--6.

[7] S. Abujar, A. K. M. Masum, M. Mohibullah, Ohidujjaman, and S. A. Hossain, ``An approach for bengali text summarization using word2vector,'' in 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1--5.

[8] M. A. I. Talukder, S. Abujar, A. K. M. Masum, F. Faisal, and S. A. Hossain, ``Bengali abstractive text summarization using sequence to sequence rnns,'' in 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1--5.

[9] M. M. Rahman, M. Aktaruzzaman Pramanik, R. Sadik, M. Roy, and P. Chakraborty, ``Bangla documents classification using transformer based deep learning models,'' in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1--5.

[10] S. Akter, A. S. Asa, M. P. Uddin, M. D. Hossain, S. K. Roy, and M. I. Afjal, ``An extractive text summarization technique for bengali document(s) using k-means clustering algorithm,'' in 2017 IEEE International Conference on Imaging, Vision Pattern Recognition (icIVPR), 2017, pp. 1--6.