Journal of Theoretical and Applied Information Technology
15th December 2019. Vol.97. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ENHANCED EXTRACTIVE TEXT SUMMARIZATION METHOD FOR MULTIPLE DOCUMENTS

[1]ADIBA MAHJABIN NITU, [2]MD. PALASH UDDIN, [3]PRIYANKA BASAK TUMPA, [4]SABINA YEASMIN, [5]MASUD IBN AFJAL

[1]Associate Professor, Hajee Mohammad Danesh Science and Technology University (HSTU), Department of Computer Science and Engineering, Dinajpur, Bangladesh

[2&5] Assistant Professor, Hajee Mohammad Danesh Science and Technology University (HSTU), Department of Computer Science and Engineering, Dinajpur, Bangladesh

[3&4] B. Sc. Student (Session 2013), Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur, Bangladesh

E-mail: [1]nitu.hstu@gmail.com, [2]palash_cse@hstu.ac.bd,[3]tumpabsk@gmail.com, [4]sabina.cse12@gmail.com,[5]masud@hstu.ac.bd

ABSTRACT

Nowadays, text summarization has become an important issue to extract the required information within short time. Several techniques on extractive text summarization have been developed for summarizing English text(s). However, there is a few works done for the summarization of Bengali text(s). In this paper, an improved extractive Bengali text summarization technique has been proposed with enhancing the word scoring process, position value heuristics and summary generation procedure of our previously presented summarizer. In the word scoring procedure, each word is preprocessed using noise removal, tokenization, stop word removal and stemming operation. Then, a heuristics is applied to calculate the word score through checking it in all the input document(s). Moreover, a modified heuristic is proposed for the sentence scoring in which it has given the priority highest to the middle sentence and then the upper and lower sentences from the middle sentence will be less prioritized. Finally, top *k*-sentences are extracted from each of the clusters of sentences made by *K*-means clustering algorithm and then the extracted sentences are sorted as their actual appearances in the original document(s). Thus, the final summary is synchronized with the original document(s). In comparison to the existing method, the experimental result shows that the proposed improved technique produces better summarization to satisfy the end-users.

Keywords: *Text Summarization, Extractive Summarization, Bengali Text Summarization, Heuristics, Synchronized Summary*

1. INTRODUCTION

The desired amount of information is collected from the massive sources such as news portals, social media, Internet transactions and so on for numerous purposes. It is often difficult to find the relevant information from the big amount of collected data [1]. Presently, the huge amount of textual data is available in many different natural languages in the era of big data which has the potentiality to be mined for extracting the exact information. Search engines such as Google, AltaVista, Yahoo etc. have been developed to retrieve the specific information from this immense amount of data. However, the outcome of a search engine is not able to provide the predicted summarized result where summary is a text that is produced from one or more texts concisely that is able to convey the important information of the original text(s). The most important benefit of using a summary is that it can reduce the studying time and monotonicity of a user. For this reason, text summarization, an application of data mining and natural language processing (NLP), is one of the most popular research areas that can allow the reader to get a quick overview of the entire document(s) [2].

Therefore, text summarization offers` an important role in the context of text mining and NLP. In this way, text summarization aims to concise the source text(s) into a shorter and precise form with preserving its information content and overall meaning [3]-[10]. In general, text summarization techniques can be classified into two categories: extractive and abstractive. Abstractive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

techniques attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of the sentences while extractive techniques perform summarization by choosing the most meaningful sentences of the document(s) according to some standard measurements on the sentences. Both techniques are used for summarizing text either for single document or multiple documents. Abstractive summarization usually needs information fusion, sentence compression and reformulation which make the abstractive summarization schemes complex because it requires deeper evaluation of source documents and concept-to-text generation as well [11]. On the other hand, extractive summarization techniques find out the most relevant sentences in the document(s) [12]. These techniques also remove the redundancy of the input text(s). Therefore, extractive summarization is less complicated than abstractive summarization to bring out the summary. However, extractive summarization involves assigning saliency measure to some units (e.g. sentences, paragraphs etc.) of the document(s) and extracting those with maximum scores to prepare the summary. The mostly used methods for extractive summarization are TF/IDF (Term Frequence/Inverse Document Frequency) sentence scoring approach, graph theoretic approach, cluster based method, machine learning based approach, query based extractive approach, regression for estimating feature weights, multilingual, topic-driven summarization, centroid-based summarization, LSA (Latent Semantic Analysis) method, networks based approach, fuzzy logic based approach etc. In this paper, an extractive summarization technique is proposed based on the TF/IDF sentence scoring approach and cluster based method for multiple Bengali documents.

Recently, there are several works done for English, Hindi and Chinese text summarization while a very few attempts have been made for Bengali text summarization. For instance, the authors of the paper described in [13] proposed a cue-based hub-authority approach for multi-document Chinese text summarization. It is a graph based summarization approach that extracts the feature words (or phrase) of various sub-topics using TF/IDF. Then, the sentence with the most ranking score within each topic is chosen as the summary. An integrated multi-document summarization approach based on hierarchical representation is presented in [14] for Chinese text. In this work, query relevancy and topic specificity are used for filtering purposes. The authors also calculated Point-wise Mutual Information (PMI) for identifying the subsumption between words and high PMI regarded as related. The authors of the paper described in [15] proposed an automatic summarization technique for Chinese text based on sub-topic partition and sentence features. In this process, the sentence weight is calculated by LexRank algorithm combining with the score of its own features such as its length, position, cue words and structure. The paper in [16] focused multi-view sentence ranking for query biased summarization of Chinese language. The authors proposed an approach that first constructs two base rankers to rank all the sentences in a document set from two independent but complementary views and then aggregates them into a consensus one. They selected the most significant content from the document set with high biased information. The paper in [17] described a sentence clustering based summarization for multi-documents text in which sentences are clustered using a similarity histogram based sentence clustering algorithm to identify multiple sub-topics (themes) from the input set of related documents and the representative sentences from the appropriate clusters are selected to form the summary. The authors of the paper in [18] presented multi-document summarization using clustering & feature specific sentence extraction. The paper [19] focused another method of multi-document summarization using sentence clustering for English language. The paper described in [20] presented text summarization using clustering technique. It is the clustering based approach that groups first, the similar documents into clusters and then sentences from every document are clustered into sentence clusters. And best scoring sentences from sentence clusters are selected to generate the final summary. For finding the similarity, cosine similarity is used. The authors in [21] proposed an extraction based summarization technique using *k*-means clustering algorithm consisting of three steps process: tokenization, computing score for each sentence and applying centroid based clustering on the sentences, and extracting important sentences as part of the summary. A multi-document text summarization for Bengali text is presented in [22] where Term Frequency (TF) based technology is used for extracting the most significant contents from a set of Bengali documents. For finding the best relevancy of sentences, the A* search algorithm is applied. Then, it finds the final summary. The paper in [23] proposed another work through sentence scoring and ranking in English. The authors of the paper explained in [24] proposed an extractive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

summarization technique for Bengali document(s) through the usage of *k*-means clustering algorithm, where they have calculated the word scores considering only the current document and given more emphasize on the first sentence for calculating the sentence score. They used TF/IFD method for calculating the scores. After that, *K*-means clustering algorithm is used to generate two clusters of the sentences for Bengali text summarization. However, the summary of the document(s) is not synchronized with the original documents. In this paper, an improved Bengali text(s) summarization is presented that uses a modified heuristics for calculating the word and sentence scores through TF/IDF method for the paper described in [24]. In addition, a proposal for the synchronization of the summary is also introduced in this paper.

The rest of this paper is formulated into the following sections. In Section 2, we describe the proposed extractive summarization technique for Bengali documents with detail explanation of the constituent steps and updated heuristics for calculating word and sentence scores. Section 3 focuses on the experimental setup and improved result analysis of different variants of the proposed summarization technique whereas Section 4 summarizes the observations and concludes the paper.

2. PROPOSED TECHNIQUE

As mentioned in the previous section, the authors of the paper described in [24] presented a technique for summarizing Bengali document(s) using *K*-means clustering algorithm [24] that lacks of synchronization specialy with a less efficient word and sentence scoring procedures. Consequently, an improvement of that Bengali summarization technique has been presented in this paper. In this proposed scheme, some heuristics for word and sentence scoring procedures along with a synchronizing technique to generate the final summary are presented. The same preprocessing steps such as noise removal, tokenization, stop word removal and stemming [24], [25] are used in this proposed approach as well. For calculating TF, each word is checked in all input documents to find its frequency. Then TF/IDF [15] is used finally to calculate the score of the word. On the other hand, the middle sentence has been given the highest priority for the sentence scoring phase and the upper and lower sentences from the middle sentence will be less prioritized. The middle sentence has been given the highest priority based on the heuristics that it can often be the most important sentence in the input document. Then, the score of each sentence is calculated by the total sum of the words' scores [23]. Moreover, if the sentence contains any cue or skeleton word, then its score is increased by 1 [23]. After that, the document is stored in a separate file with the sentences' scores.

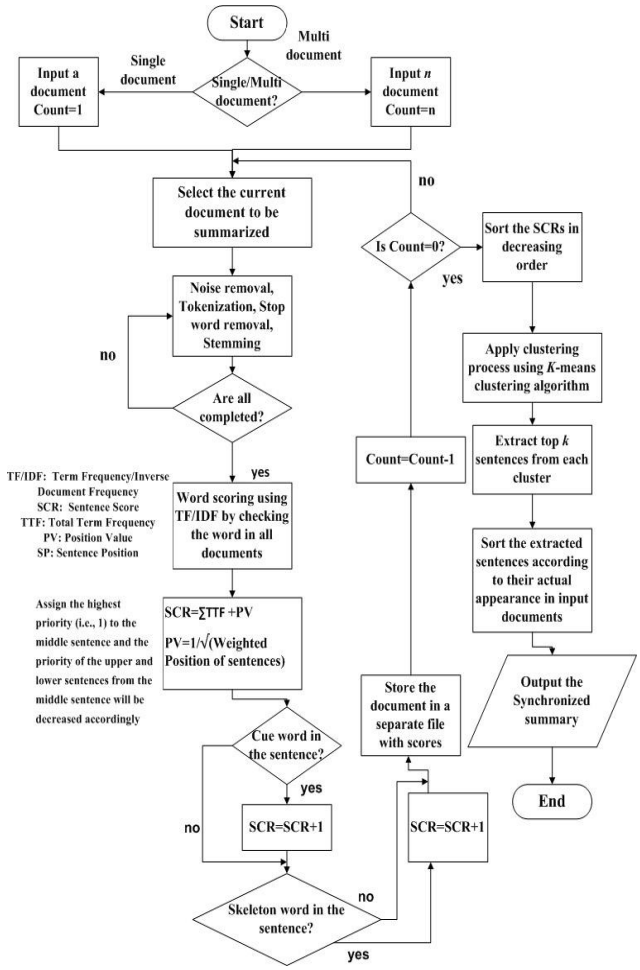

Figure 1: Flowchart of improved extractive Bengali document(s) summarization technique

For multiple documents summarization, the preprocessing, word scoring and sentence scoring operations are repeated as mentioned above and also the results are placed into the same file in such a way that all the documents are merged. Subsequently, the sentences are sorted in decreasing order of scores and it has been taken into consideration that the highest score is as centroid 1 and the lowest score is as centroid 2 to perform the *K*-means clustering algorithm [21] [26] with a view to creating the two clusters of sentences. After getting the final two clusters of sentences using the *K*-means clustering algorithm, top *k* sentences from each cluster are extracted to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

get the summary. Here, k sentences can be measured as 10-40% of total sentences of the original document(s). Finally, the sentences in the generated summary are sorted as their actual appearances in the original document(s) to offer the synchronization with the original input document(s). The discussed working steps of the proposed improved summarization technique are illustrated in the following flowchart of Fig. 1.

The proposed extractive text summarization technique is further discussed here in step-by-step:

2.1 Preprocessing

The operations performed in this step are noise removal, tokenization, stemming and stop word removal. Noise removal is involved with removing header, footer, etc., from the document. Tokenization separates each word into lexical form. Words are separated with the aid of কমা, দাঁড়ি etc. A word can be found in different forms in the same document. Stemming means the Bengali words need to be converted to their original form for simplicity. The stemming algorithm is used to transform words to their canonical forms, like অনুষ্ঠানকে, অনুষ্ঠানের, অনুষ্ঠানেও, অনুষ্ঠানে etc. should be converted to their original form অনুষ্ঠান. The rules for stemming any word are discussed in [24], [25] in detail. In the proposed technique, the following rules for stemming any Bengali word are used:

i. X # When X appears at the end of a word, remove it.
ii. Y → Z # When Y appears at the end of a word, replace it with Z
iii. Y.Z → A.B # When Y, followed by some character a, followed by Z or # appears at the end of a word, replace with AaB

Table 1: Examples of Bengali word stemming.

Suffix	Original words	After Stemming
ই	# যেখানেই, সেখানেই	#যেখানে,সেখানে
তো	# নয়তো, করতো	# নয়,কর
কে	# সেটাকে, তাহাকে	# সেটা,তাহা
ে ে-> া	# ভেসে, কেঁদে	# ভাসা,কাঁদা
ে ে ছিলেন →া	# কেঁদেছিলেন, ভেসেছিলেন	# কাঁদা, ভাসা

Let's see an example: সাজিদ কাজ করছে. After stemming, it will be সাজিদ কাজ কর. Some examples ar eshown in Table 1. Consider the words #যেখানেই, সেখানেই since ই, কে, তো appear at the end of any word, they are removed from the original word (according to the rule#01). For #মেতে, যেচে since ে, ে appear at the end of a word, they are replaced with া (according to the rule#02). Again for #কেঁদেছিলেন, ভেসেছিলেন since ে, ে ছিলেন appear at the end of a word, they are replaced with □ and will be #কাঁদা, ভাসা (according to the rule#03).

2.2 Scoring Procedure

For word score calculation, TF/IDF approach is used. If there are more unique words in a given sentence, then the sentence is relatively more important [23]. In this technique, the number of times that each word occurs in all documents is considered for the calculation of its score. The TF/IDF to calculate the word score is defined as follows:

$$TF = tfw_{i,s} * idfw_i \quad (1)$$

$$idfw_i = \log\left(\frac{TotSen}{n_i}+1\right) \quad (2)$$

Where, TF=Term Frequency, $tfw_{i,s}$ =Number of occurrences of the word i in the sentence S, $idfw_i$ =Inverse document frequency, N=Total number of the sentences in the text and n_i =Number of sentences of all documents in which word w_i occurs. Now, the sentence score (SC) is found as:

$$SC_{CHK} = \sum TF + PV \quad (3)$$

$$PV = \frac{1}{\sqrt{Sp}} \quad (4)$$

Here, Sp = Position of the sentences and PV =Position value. In this case of PV, we have updated the existing assignment mechanism. In the proposed technique, we have assigned the highest priority (i. e., 1) to the middle sentence and the priority of the upper and lower sentences from the middle sentence will be decreased accordingly.

2.3 Generation of Synchronized Summary

Calculating the entire sentences' scores K-means clustering algorithm is applied. For this, the sentences' scores are sorted. After that, the lowest and the highest scores are assigned to the two clusters' centroid M1 and M2, and the distance from each centroid to each sentence is measured. Nearest distance closest to one centroid means that the sentence belongs to that cluster. Thus, two clusters are created and for next iteration, centroid values are updated. For updating the centroid

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

values, the average values of each cluster are calculated and assigned them as new centroids. This process is repeated until two consecutive iterations produce the same result. Then, top *k* sentences are extracted from each cluster to generate the summary. Moreover, extracted *k*-sentences from each cluster are sorted according to their actual appearance in the original documents. In t his way, the generated summary is synchronized with the original document(s).

3. EXPERIMENT AND RESULT ANALYSIS

For experimental result analysis, let us consider two speech documents to be summarized. Among which the first document can be found at [27] while the second document can be found at [27] which are also place below.

1st Document:

মাননীয় অধ্যক্ষ মহোদয়, শ্রদ্ধেয় শিক্ষকমণ্ডলী, অভিভাবকবৃন্দ এবং প্রীতিভাজন নবাগত ও সতীর্থ শিক্ষার্থী ভাইবোনসবার প্রতি রইল — আমার সালাম ও শুভেচ্ছা।আমাদের এ ঐতিহ্যবাহী শিক্ষাপ্রতিষ্ঠান আজ নতুন প্রাণের স্পর্শে স্পন্দিত ও আনন্দ হিল্লোলে মুখরিত। নবাগত শিক্ষার্থী ভাই ও বোনেরা, স্কুলের গণ্ডি পেরিয়ে তোমরা এসেছ আমাদের কলেজের আঙিনায়, ছোট্ট গণ্ডি পেরিয়ে মুক্ত জ্ঞানের আলোয়, উচ্চতর শিক্ষা অর্জনের প্রাথমিক ক্ষেত্রে। তোমাদের সবার প্রতি রইল আমার আন্তরিক উষ্ণ অভিনন্দন ও শুভেচ্ছা।নবাগত ভাই ও বোনেরা,তোমরা জেনে খুশি হবে যে এই অঞ্চলের মধ্যে আমাদের এ প্রতিষ্ঠানটি নানা দিক থেকে গৌরবোজ্জ্বল ঐতিহ্যের অধিকারী। এক মহান ব্রত ধারণ করে আছে এ পবিত্র বিদ্যাঙ্গণ। যার স্লোগান হচ্ছেশিক্ষাই শক্তি। শিক্ষার জন্য এসো‘ —, সেবার জন্য বেরিয়ে যাও।আজ থেকে তোমাদের জন্য এটাই হোক জীবনের মূলমন্ত্র। ’ র জন্য তোমাদের সে অভিযাত্রা শুভ হোকআলোকিত জীবনে, সুন্দর হোক।বন্ধুরা,তোমরা জেনে খুুশি হবে যে আমাদের এ কলেজ থেকেই অতীতে পড়ালেখা করে জ্ঞানীগুণী পণ্ডিত হিসেবে দেশে ও আন্তর্জাতিক ক্ষেত্রে সুনাম অর্জন করেছেন অনেকে। ডাক্তার, প্রকৌশলী, রাজনীতিক, সাহিত্যিক, সাংবাদিক হিসেবে এ প্রতিষ্ঠানের অনেক শিক্ষার্থীই পরবর্তী জীবনে প্রতিষ্ঠা লাভ করেছেন সমাজে। তোমরাও তাদের মতোই দেশ ও জাতির মুখ উজ্জ্বল করবে।নবীন বন্ধুরা,আমাদের এ শিক্ষালয়ের দীর্ঘদিনের এক সুমহান ঐতিহ্য ও খ্যাতি রয়েছে দেশব্যাপী। এখানকার শিক্ষাশৃঙ্খলা-, ছাত্রশিক্ষক সুসম্পর্ক-, পরীক্ষায় ঈর্ষণীয় ফলাফল সারাদেশে আলোড়ন সৃষ্টি করে প্রতিবছরই। আজ তোমরা যারা নতুন প্রাণশক্তি ও অমিত সম্ভাবনা নিয়ে এখানে এসেছ, তোমাদের সাদরে বরণ করে নিয়ে তোমাদেরকেও ঐতিহ্য রক্ষার সুমহান দায়িত্বে অংশীদার করছি। আশা করি, এ প্রতিষ্ঠানের গৌরব বৃদ্ধিতে তোমরাও প্রাণপণ চেষ্টা করবে।বন্ধুরা,এ প্রতিষ্ঠানে তোমাদের শিক্ষাজীবন সফলসার্থক ও গৌরবময় হোক। এখানকার শিক্ষা নিয়ে তোমরা -জাতি ও বিশ্বমানবতার কল্যাণে নিয়োজিত হও। যৌবনের আলোয় -দেশ আলোকিত হোক তোমাদের ভুবনএই প্রত্যাশায় শেষ করছি —। সবাইকে ধন্যবাদ, সবার জন্য শুভকামনা।

2nd Document:

মাননীয় সভাপতি ও বরেণ্য অতিথিবৃন্দ। আজকের এই মহতী ও গুরুত্বপূর্ণ আলোচনা সভায় আমাকে প্রধান অতিথি হিসেবে সম্মানিত করার জন্য আমি আয়োজকদের জানাই কৃতজ্ঞতা।সুধিবৃন্দ,যে জাতি শিক্ষিত নয়, সেই জাতি উন্নতির শিখরে আরোহণ করতে বাধাগ্রস্ত হয়। আজকের শিশু আগামী দিনের ভবিষ্যত্। তারা যদি অনগ্রসর থাকে তাহলে জাতি তার গতি হারিয়ে ফেলবে। তাই জাতিকে উন্নতির চরম শিখরে আরোহণ করতে হলে সেই জাতির প্রত্যেক ব্যক্তিকে শিক্ষার আলোয় আলোকিত এবং স্বশিক্ষিত হতে হবে।প্রিয় শিক্ষা অনুরাগী,আলোকিত মানুষ দেশপ্রেম, ধর্মীয় মূল্যবোধ, শৃঙ্খলা, শিষ্টাচার, শ্রমের মর্যাদা এবং বেকারত্বের অভিশাপ থেকে দেশকে মুক্ত করে দেশকে উন্নতির পথ দেখাতে পারে।আলোকিত মানুষ গড়ার কাজটি খুব সহজ নয়। ঠিক এ মুহূর্তে একবার ভেবে দেখুন আলোকিত মানুষের অভাবে আমরা আজ কত নিচে আছি। শুধু প্রাতিষ্ঠানিক শিক্ষাই যথার্থ শিক্ষা নয়। এর সাথে স্বশিক্ষায় জনগণকে শিক্ষিত হতে হবে। আর আজকের এই প্রচেষ্টা শুধু একজনকে করলে অথবা এই অনুষ্ঠানের মাধ্যমে শেষ হয়ে গেলে চলবে না, এর জন্য সরকার ও জনগণ সবাইকে এগিয়ে আসতে হবে। সবাইকে এক সাথে উচ্চারণ করতে হবে“— চলো যাই পড়তে যাই, আলোকিত মানুষ চাই ।”সবার সুস্থ, সুন্দর জীবন কামনা করে, সবাইকে ধন্যবাদ জানিয়ে আমি আমার বক্তব্য শেষ করছি। ধন্যবাদ।

Getting the sentences' score of all documents using the improved scoring process, the scores with sentences have been stored in a merged file which is shown in Table 2. Then, *K*-means clustering algorithm has been applied with initial centroids as M1=54.40 and M2=2.49. As discussed earlier, the final iteration of the clustering algorithm is shown in Table 3.

Table 2: Merged file with sentence scores.

Sentence Number	Score	Sentence
(1)	54.40	নবাগত শিক্ষার্থী ভাই ও বোনেরা, স্কুলের গণ্ডি পেরিয়ে তোমরা এসেছ আমাদের কলেজের আঙিনায়, ছোট্ট গণ্ডি পেরিয়ে মুক্ত জ্ঞানের আলোয়, উচ্চতর শিক্ষা অর্জনের প্রাথমিক ক্ষেত্রে।
(2)	51.63	প্রিয় শিক্ষা অনুরাগী, আলোকিত মানুষ দেশপ্রেম, ধর্মীয় মূল্যবোধ, শৃঙ্খলা, শিষ্টাচার, শ্রমের মর্যাদা এবং বেকারত্বের অভিশাপ থেকে দেশকে মুক্ত করে দেশকে উন্নতির পথ দেখাতে পারে।
(3)	47.94	আর আজকের এই প্রচেষ্টা শুধু একজনকে করলে অথবা এই অনুষ্ঠানের মাধ্যমে শেষ হয়ে গেলে চলবে না, এর জন্য সরকার ও জনগণ সবাইকে এগিয়ে আসতে হবে।
(4)	47.54	আজ তোমরা যারা নতুন প্রাণশক্তি ও অমিত সম্ভাবনা নিয়ে এখানে এসেছ, তোমাদের সাদরে বরণ করে নিয়ে তোমাদেরকেও ঐতিহ্য রক্ষার সুমহান দায়িত্বে অংশীদার করছি।
(5)	43.42	বন্ধুরা, তোমরা জেনে খুশি হবে যে আমাদের এ কলেজ থেকেই অতীতে

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		পড়ালেখা করে জ্ঞানীগুণী পণ্ডিত হিসেবে দেশে ও আন্তর্জাতিক ক্ষেত্রে সুনাম অর্জন করেছেন অনেকে।
(6)	42.25	মাননীয় অধ্যক্ষ মহোদয়, শ্রদ্ধেয় শিক্ষকমণ্ডলী, অভিভাবকবৃন্দ এবং প্রীতিভাজন নবাগত ও সতীর্থ শিক্ষার্থী ভাইবোন— সবার প্রতি রইল আমার সালাম ও শুভেচ্ছা।
(7)	36.30	নবাগত ভাই ও বোনেরা, তোমরা জেনে খুশি হবে যে এই অঞ্চলের মধ্যে আমাদের এ প্রতিষ্ঠানটি নানা দিক থেকে গৌরবোজ্জ্বল ঐতিহ্যের অধিকারী।
(8)	35.95	তাই জাতিকে উন্নতির চরম শিখরে আরোহণ করতে হলে সেই জাতির প্রত্যেক ব্যক্তিকে শিক্ষার আলোয় আলোকিত এবং স্বশিক্ষিত হতে হবে।
(9)	32.21	আজকের এই মহতী ও গুরুত্বপূর্ণ আলোচনা সভায় আমাকে প্রধান অতিথি হিসেবে সম্মানিত করার জন্য আমি আয়োজকদের জানাই কৃতজ্ঞতা।
(10)	31.64	ডাক্তার, প্রকৌশলী, রাজনীতিক, সাহিত্যিক, সাংবাদিক হিসেবে এ প্রতিষ্ঠানের অনেক শিক্ষার্থীই পরবর্তী জীবনে প্রতিষ্ঠা লাভ করেছেন সমাজে।
(11)	28.60	সুধিবৃন্দ, যে জাতি শিক্ষিত নয়, সেই জাতি উন্নতির শিখরে আরোহণ করতে বাধাগ্রস্ত হয়।
(12)	27.10	ঠিক এ মুহূর্তে একবার ভেবে দেখুন আলোকিত মানুষের অভাবে আমরা আজ কত নিচে আছি।
(13)	26.84	সবাইকে এক সাথে উচ্চারণ করতে হবে— " চলো যাই পড়তে যাই, আলোকিত মানুষ চাই।
(14)	26.28	এখানকার শিক্ষা-শৃঙ্খলা, ছাত্র-শিক্ষক সুসম্পর্ক, পরীক্ষায় ঈর্ষণীয় ফলাফল সারাদেশে আলোড়ন সৃষ্টি করে প্রতিবছরই।
(15)	25.89	সবার সুস্থ, সুন্দর জীবন কামনা করে, সবাইকে ধন্যবাদ জানিয়ে আমি আমার বক্তব্য শেষ করছি।
(16)	24.96	আমাদের এ ঐতিহ্যবাহী শিক্ষাপ্রতিষ্ঠান আজ নতুন প্রাণের স্পর্শে স্পন্দিত ও আনন্দ হিল্লোলে মুখরিত।
(17)	23.42	নবীন বন্ধুরা, আমাদের এ শিক্ষালয়ের দীর্ঘদিনের এক সুমহান ঐতিহ্য ও খ্যাতি রয়েছে দেশব্যাপী।
(18)	22.16	তারা যদি অনগ্রসর থাকে তাহলে জাতি তার গতি হারিয়ে ফেলবে।
(19)	20.04	আলোকিত জীবনের জন্য তোমাদের সে অভিযাত্রা শুভ হোক, সুন্দর হোক।
(20)	19.02	আশা করি, এ প্রতিষ্ঠানের গৌরব বৃদ্ধিতে তোমরাও প্রাণপণ চেষ্টা করবে।
(21)	18.45	এখানকার শিক্ষা নিয়ে তোমরা দেশ-জাতি ও বিশ্বমানবতার কল্যাণে নিয়োজিত হও।
(22)	17.75	এক মহান ব্রত ধারণ করে আছে এ পবিত্র বিদ্যাঙ্গণ।
(23)	17.70	তোমাদের সবার প্রতি রইল আমার আন্তরিক উষ্ণ অভিনন্দন ও শুভেচ্ছা।
(24)	17.64	তোমরাও তাদের মতোই দেশ ও জাতির মুখ উজ্জ্বল করবে।
(25)	16.74	যৌবনের আলোয় আলোকিত হোক তোমাদের ভুবন— এই প্রত্যাশায় শেষ করছি।
(26)	16.41	শিক্ষার জন্য এসো, সেবার জন্য বেরিয়ে যাও।
(27)	15.08	শুধু প্রাতিষ্ঠানিক শিক্ষাই যথার্থ শিক্ষা নয়।
(28)	14.34	এর সাথে স্বশিক্ষায় জনগণকে শিক্ষিত হতে হবে।
(29)	13.88	বন্ধুরা, এ প্রতিষ্ঠানে তোমাদের শিক্ষাজীবন সফল-সার্থক ও গৌরবময় হোক।
(30)	13.58	আলোকিত মানুষ গড়ার কাজটি খুব সহজ নয়।
(31)	12.20	মাননীয় সভাপতি ও বরেণ্য অতিথিবৃন্দ।
(32)	11.62	আজ থেকে তোমাদের জন্য এটাই হোক জীবনের মূলমন্ত্র।
(33)	11.58	আজকের শিশু আগামী দিনের ভবিষ্যত্।
(34)	11.34	যার স্লোগান হচ্ছে— 'শিক্ষাই শক্তি।
(35)	9.59	সবাইকে ধন্যবাদ, সবার জন্য শুভকামনা।
(36)	2.49	ধন্যবাদ।

Table 3: Final iteration.

Centroid	Cluster no.	Score	Sentence
M1=54.46	Cluster-1	54.40	নবাগত শিক্ষার্থী ভাই ও বোনেরা, স্কুলের গণ্ডি পেরিয়ে তোমরা এসেছ আমাদের কলেজের আঙিনায়, ছোট্ট গণ্ডি পেরিয়ে মুক্ত জ্ঞানের আলোয়, উচ্চতর শিক্ষা অর্জনের প্রাথমিক ক্ষেত্রে।
		51.63	প্রিয় শিক্ষা অনুরাগী,আলোকিত মানুষ দেশপ্রেম, ধর্মীয় মূল্যবোধ, শৃঙ্খলা, শিষ্টাচার, শ্রমের মর্যাদা এবং বেকারত্বের অভিশাপ থেকে দেশকে মুক্ত করে দেশকে উন্নতির পথ দেখাতে পারে।
		47.94	আর আজকের এই প্রচেষ্টা শুধু একজনকে করলে অথবা এই অনুষ্ঠানের মাধ্যমে শেষ হয়ে গেলে চলবে না, এর জন্য সরকার ও জনগণ

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

			সবাইকে এগিয়ে আসতে হবে।
		47.54	আজ তোমরা যারা নতুন প্রাণশক্তি ও অমিত সম্ভাবনা নিয়ে এখানে এসেছ, তোমাদের সাদরে বরণ করে নিয়ে তোমাদেরকেও ঐতিহ্য রক্ষার সুমহান দায়িত্বে অংশীদার করছি।
		43.42	বন্ধুরা, তোমরা জেনে খুশি হবে যে আমাদের এ কলেজ থেকেই অতীতে পড়ালেখা করে জ্ঞানীগুণী পণ্ডিত হিসেবে দেশে ও আন্তর্জাতিক ক্ষেত্রে সুনাম অর্জন করেছেন অনেকে।
		42.25	মাননীয় অধ্যক্ষ মহোদয়, শ্রদ্ধেয় শিক্ষকমণ্ডলী, অভিভাবকবৃন্দ এবং প্রীতিভাজন নবাগত ও সতীর্থ শিক্ষার্থী ভাইবোন— সবার প্রতি রইল আমার সালাম ও শুভেচ্ছা।
		36.30	নবাগত ভাই ও বোনেরা, তোমরা জেনে খুশি হবে যে এই অঞ্চলের মধ্যে আমাদের এ প্রতিষ্ঠানটি নানা দিক থেকে গৌরবোজ্জ্বল ঐতিহ্যের অধিকারী।
		35.95	তাই জাতিকে উন্নতির চরম শিখরে আরোহণ করতে হলে সেই জাতির প্রত্যেক ব্যক্তিকে শিক্ষার আলোয় আলোকিত এবং স্বশিক্ষিত হতে হবে।
		32.22	আজকের এই মহতী ও গুরুত্বপূর্ণ আলোচনা সভায় আমাকে প্রধান অতিথি হিসেবে সম্মানিত করার জন্য আমি আয়োজকদের জানাই কৃতজ্ঞতা।
		31.64	ডাক্তার, প্রকৌশলী, রাজনীতিক, সাহিত্যিক, সাংবাদিক হিসেবে এ প্রতিষ্ঠানের অনেক শিক্ষার্থীই পরবর্তী জীবনে প্রতিষ্ঠা লাভ করেছেন সমাজে।
		28.60	সুধিবৃন্দ, যে জাতি শিক্ষিত নয়, সেই জাতি উন্নতির শিখরে আরোহণ করতে বাধাগ্রস্ত হয়।
		27.19	ঠিক এ মুহূর্তে একবার ভেবে দেখুন আলোকিত মানুষের অভাবে আমরা আজ কত

			নিচে আছি।
		26.84	সবাইকে এক সাথে উচ্চারণ করতে হবে—" চলো যাই পড়তে যাই, আলোকিত মানুষ চাই।
		26.28	এখানকার শিক্ষা-শৃঙ্খলা, ছাত্র-শিক্ষক সুসম্পর্ক, পরীক্ষায় ঈর্ষণীয় ফলাফল সারাদেশে আলোড়ন সৃষ্টি করে প্রতিবছরই।
		25.89	" সবার সুস্থ, সুন্দর জীবন কামনা করে, সবাইকে ধন্যবাদ জানিয়ে আমি আমার বক্তব্য শেষ করছি।
		24.96	আমাদের এ ঐতিহ্যবাহী শিক্ষাপ্রতিষ্ঠান আজ নতুন প্রাণের স্পর্শে স্পন্দিত ও আনন্দ হিল্লোলে মুখরিত।
		23.42	নবীন বন্ধুরা, আমাদের এ শিক্ষালয়ের দীর্ঘদিনের এক সুমহান ঐতিহ্য ও খ্যাতি রয়েছে দেশব্যাপী।
		22.16	তারা যদি অনগ্রসর থাকে তাহলে জাতি তার গতি হারিয়ে ফেলবে।
		20.04	আলোকিত জীবনের জন্য তোমাদের সে অভিযাত্রা শুভ হোক, সুন্দর হোক।
		19.02	আশা করি, এ প্রতিষ্ঠানের গৌরব বৃদ্ধিতে তোমরাও প্রাণপণ চেষ্টা করবে।
		18.45	এখানকার শিক্ষা নিয়ে তোমরা দেশ-জাতি ও বিশ্বমানবতার কল্যাণে নিয়োজিত হও।
		17.75	এক মহান ব্রত ধারণ করে আছে এ পবিত্র বিদ্যাঙ্গণ।
		17.70	তোমাদের সবার প্রতি রইল আমার আন্তরিক উষ্ণ অভিনন্দন ও শুভেচ্ছা।
M2=27.43	Cluster-2	17.64	তোমরাও তাদের মতোই দেশ ও জাতির মুখ উজ্জ্বল করবে।
		16.74	যৌবনের আলোয় আলোকিত হোক তোমাদের ভুবন— এই প্রত্যাশায় শেষ করছি।
		16.41	শিক্ষার জন্য এসো, সেবার জন্য বেরিয়ে যাও।
		15.08	শুধু প্রাতিষ্ঠানিক শিক্ষাই যথার্থ শিক্ষা নয়।
		14.34	এর সাথে স্বশিক্ষায় জনগণকে শিক্ষিত হতে হবে।

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		13.88	বন্ধুরা, এ প্রতিষ্ঠানে তোমাদের শিক্ষাজীবন সফল-সার্থক ও গৌরবময় হোক।
		13.58	আলোকিত মানুষ গড়ার কাজটি খুব সহজ নয়।
		12.20	মাননীয় সভাপতি ও বরেণ্য অতিথিবৃন্দ।
		11.62	আজ থেকে তোমাদের জন্য এটাই হোক জীবনের মূলমন্ত্র।
		11.58	আজকের শিশু আগামী দিনের ভবিষ্যত্।
		11.34	যার স্লোগান হচ্ছে— 'শিক্ষাই শক্তি।
		9.59	সবাইকে ধন্যবাদ, সবার জন্য শুভকামনা।
		2.49	ধন্যবাদ।

To this end, after extracting top 5 (*k*) sentences from each cluster, the sentences are sorted according to their actual appearance in the original input documents. The finally synchronized summary generated using the IDE for Java application, "Netbeans IDE 8.0", running on a 2.50 GHz Intel® core™ *i*5 CPU having 4GB RAM with Windows 7 ultimate operating system, which contains 10 sentences, is shown below:

নবাগত শিক্ষার্থী ভাই ও বোনেরা, স্কুলের গণ্ডি পেরিয়ে তোমরা এসেছ আমাদের কলেজের আঙিনায়, ছোট্ট গণ্ডি পেরিয়ে মুক্ত জ্ঞানের আলোয়, উচ্চতর শিক্ষা অর্জনের প্রাথমিক ক্ষেত্রে। প্রিয় শিক্ষা অনুরাগী, আলোকিত মানুষ দেশপ্রেম, ধর্মীয় মূল্যবোধ, শৃঙ্খলা, শিষ্টাচার, শ্রমের মর্যাদা এবং বেকারত্বের অভিশাপ থেকে দেশকে মুক্ত করে দেশকে উন্নতির পথ দেখাতে পারে । বন্ধুরা, তোমরা জেনে খুশি হবে যে আমাদের এ কলেজ থেকেই অতীতে পড়ালেখা করে জ্ঞানীগুণী পণ্ডিত হিসেবে দেশে ও আন্তর্জাতিক ক্ষেত্রে সুনাম অর্জন করেছেন অনেকে। আর আজকের এই প্রচেষ্টা শুধু একজনকে করলে অথবা এই অনুষ্ঠানের মাধ্যমে শেষ হয়ে গেলে চলবে না, এর জন্য সরকার ও জনগণ সবাইকে এগিয়ে আসতে হবে। আজ তোমরা যারা নতুন প্রাণশক্তি ও অমিত সম্ভাবনা নিয়ে এখানে এসেছ, তোমাদের সাদরে বরণ করে নিয়ে তোমাদেরকেও ঐতিহ্য রক্ষার সুমহান দায়িত্বে অংশীদার করছি। সুধিবৃন্দ, যে জাতি শিক্ষিত নয়, সেই জাতি উন্নতির শিখরে আরোহণ করতে বাধাগ্রস্ত হয়। ঠিক এ মুহূর্তে একবার ভেবে দেখুন আলোকিত মানুষের অভাবে আমরা আজ কত নিচে আছি। সবাইকে এক সাথে উচ্চারণ করতে হবে—" চলো যাই পড়তে যাই, আলোকিত মানুষ চাই।" সবার সুস্থ, সুন্দর জীবন কামনা করে, সবাইকে ধন্যবাদ জানিয়ে আমি আমার বক্তব্য শেষ করছি। এখানকার শিক্ষা-শৃঙ্খলা, ছাত্র-শিক্ষক সুসম্পর্ক, পরীক্ষায় ঈর্ষণীয় ফলাফল সারাদেশে আলোড়ন সৃষ্টি করে প্রতিবছরই।

However, the summary of the two documents using the existing summarization technique can be found at [24]. It is clearly seen that the summary of the same two documents produced by the improved summarization technique can satisfy the end-users more than the summary produced by the existing technique. Thus, the summary produce by the proposed technique can carry more meanings than that of the existing technique.

3.1 Comparative Discussion

The presented extractive Bengali text summarization is implemented in several ways through varying the proposed parameters for word and sentence scoring on its preceding technique to measure the effectiveness. In the variants of the proposed method, the newly added aforementioned two heuristics (one is called position value heuristics and the other is for word scoring heuristics) and the synchronization scheme are taken or not for the different combinations as shown in Table 4. As an instance for variant 6, the position value heuristics for sentence scoring and the synchronization strategy have been updated in comparson to its preceeding existing method while the word scoring heuristics is remained unchanged.

Table 4: Variants of the proposed technique.

Position value Heuristics	Word Score Heuristics	Synchronization scheme	Variation of Proposed Method (PM)
No (unchanged)	No	No	Original
Yes (changed)	No	No	Variant 1
No	Yes	No	Variant 2
No	No	Yes	Variant 3
Yes	Yes	No	Variant 4
No	Yes	Yes	Variant 5
Yes	No	Yes	Variant 6
Yes	Yes	Yes	Variant 7

To analyze the effectiveness of the variants of the proposed method over the existing method, the two Bengali documents from [28], [29] have been considered which are used in the preceding technique [24] of the proposed method. Now, the summaries have been produced by the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

original method and the different variants of the proposed technique. From the produced results of the summarization technique, the variants of the proposed technique can strikingly produce different summary than the original method. Morever, the summary produced by the variant 7 can carry more meanings than that of the others variants satisfying the users rigorously. However, a questionnaire has been conducted among different 100 Bengali native people to get the feedback on the different summaries produced by the variants of the presented method. For validation of the proposed method, the feedback (average satisfactory rate on the summary) on the 8 different summaries for the considered input documents is shown in the following graph:

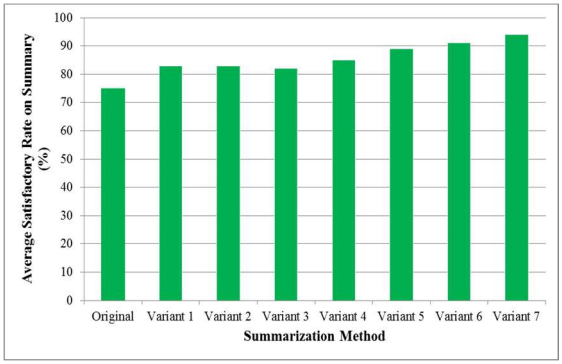

Fig. 2. End-user feedback (average satisfactory rate on the summaries)) by the existing and proposed methods

In this way, from the above graph it can also be observed that the summary variant 7 can satisfy the user more (94% in average) than the others. In another words, it has been established that the proposed two heuristics for calculating word and sentence scores and synchronized summary make-up procedure for Bengali documents can successfully be employed for real-life summary generation.

4. CONCLUSION

In this paper, an improved extraction based Bengali text summarization technique is presented for both single and multiple documents. The proposed extractive text summarization provides the summary where some more significant sentences are extracted from the original document(s). The effectiveness of the proposed technique is improved in comparison to the existing solutions because of using two heuristics for the word and sentence scoring processes along with a new technique for generating synchronized summary. In particular, *K*-means algorithm is used for making groups of related sentences. Then, from the group of sentences, the top-most meaningful sentences are selected and they are sorted as their actual appearance in the original document(s). In this way, the final summary is synchronized with the original document(s). The result is compared with existing technique and obtained the better users' satisfaction on the generated summary. According to the result of the proposed technique, it can be concluded that it reduces the redundancy and provides better summarization of Bengali text document(s).

REFRENCES:

[1] E. Turban, J. E. Aronson, T.-P. Liang and R. Sharda, *Decision Support and Business Intelligence Systems*. 8th Edition, Prentice-Hall, Inc., USA 2006.

[2] S. A. Babar, "Text Summarization: An Overview", *Technical Notes*, M.Tech-CSE, RIT.

[3] H. Dave and S. Jaswal, "Multiple Text Document Summarization System using Hybrid Summarization Technique", *I st International Conference on Next Generation Computing Technologies*, pp. 804-808, September 2015, India.

[4] A. Khan and N. Salim, "A Review on Abstractive Summarization Methods", *Journal of Theoretical and Applied Information Technology*, Vol. 59, No.1, pp. 64-72, 2014.

[5] S. Patil and R. Joshi, "Enrich Framework for Multi-Document Summarization Using Text Features and Fuzzy Logic", *Journal of Theoretical and Applied Information Technology*, Vol.88. No.3, pp. 431-437, 2016.

[6] A. Elrefaiy, A. R. Abas and I. Elhenawy, "Review of Recent Techniques for Extractive Text Summarization", *Journal of Theoretical and Applied Information Technology*, Vol. 96. No. 23, pp. 7739-7759, 2018.

[7] S. Jusoh, "A Study on NLP Applications and Ambiguity Problems", *Journal of Theoretical and Applied Information Technology*, Vol. 96. No. 6, pp. 1486-1499, 2018.

[8] P. B. Tumpa, S. Yeasmin, A. M. Nitu, M. P. Uddin, M. I. Afjal and M. A. A. Mamun, "An Improved Extractive Summarization Technique for Bengali Text(s)", *IEEE International Conference on Computer, Communication, Chemical, Material and*

Electronic Engineering (IC4ME2), 8-9 February 2018, Bangladesh.

[9] S. Yeasmin, P. B. Tumpa, A. M. Nitu, M. P. Uddin, E. Ali and M. I. Afjal, "Study of Abstractive Text Summarization Techniques", *American Journal of Engineering Research (AJER)*, Vol. 06, Issue. 8, pp. 253-260, 2017.

[10] A. S. Asa, S. Akter, M. P. Uddin, M. D. Hossain, S. K. Roy and M. I. Afjal, "A Comprehensive Survey on Extractive Text Summarization Techniques", *American Journal of Engineering Research (AJER)*, Vol. 06, Issue. 1, pp. 200-205, 2017.

[11] N. R. Kasture, N. Yargal, N. Singh, N. Kulkarni and V. Mathur, "A Survey on Methods of Abstractive Text Summarization", *International Journal for Research in Emerging Science and Technology*, Vol. 1, Issue. 6, 2014.

[12] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques", *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, pp. 258-268, 2010.

[13] J. Zhang, L. Sun and Q. Zhou, "Cue-based Hub-Authority approach for Multi- document Text Summarization", *International Conference on Natural Language Processing and Knowledge Engineering*, pp. 642 – 645, 2005, Beijing.

[14] Y. Ouyang, W. Li and Q. Lu , "An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation" , *ACLShort '09 Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 113-116, 2009, China.

[15] X. Li, J. Zhang and M. Xing, "Automatic Summarization for Chinese text based on Sub Topic Partition and Sentence Features*", IEEE 2nd International Symposium on Intelligence Information Processing and Trusted Computing*, 22-23 October 2011, China.

[16] P. Hu, T. He and H. Wang, "Multi-View Sentence Ranking for Query-Biased Summarization", *International Conference on Computational Intelligence and Software Engineering (CiSE)*, 2010, China.

[17] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents" *TECHNIA–International Journal of Computing Science and Communication Technologies*, Vol. 2, No. 1, pp. 325-335, 2009.

[18] A. Kogilavani, and P. Balasubramani, "Clustering and Feature specific sentence extraction based summarization of multidocuments*", International Journal of Computer Science & Information Technology,* Vol.2, No.4, pp. 99-111, 2010.

[19] T. J. Siddiqui and V. K. Gupta, "Multi-document Summarization using Sentence Clustering", *IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction*, 27-29 December 2012, India.

[20] A. R. Deshpande and L. M. R. J. Lobo, "Text Summarization using Clustering Technique", *International Journal of Engineering Trends and Technology,*Vol. 4, Issue. 8, pp. 3348-3351, 2013.

[21] A. Agrawal and U. Gupta, "Extraction based approach for text summarization using k-means clustering", *International Journal of Scientific and Research Publications*, Vol. 4, Issue. 11, pp. 1-4, 2014.

[22] M. A. Uddin, K. Z. Sultana and M. A. Alom, "A Multi-Document Text Summarization for Bengali Text*", IEEE International Forum on Strategic Technology (IFOST),* 2014, Bangladesh.

[23] M. I. A. Efat, M. Ibrahim and H. Kayesh, "Automated Bangla Text Summarization by Sentence Scoring and Ranking", *IEEE International Conference on Informatics, Electronics & Vision (ICIEV)*, 2013, Bangladesh.

[24] 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", *IEEE International Conference on imaging, vision & pattern recognition (icVPR)*, 2017, Bangladesh.

[25] R. Kamal, "Bangla-Stemmer", [Online]. Available: https://github.com/rafi-kamal/Bangla-Stemmer, 14 Nov 2018.

[26] A. Mhatre, "Implementation of K-Means Algorithm in C++", [Online]. Available: http://ankurm.com/implementation-of-k-means-algorithm-in-c/, 15 May 2018.

[27] □□. □□□□□□ □□□□, "নবাগত ছাত্রছাত্রীদের উদ্দেশ্যে একটি ভাষণ", [Online]. Available: https://archive1.ittefaq.com.bd/print-edition/aunoshilon/2015/05/17/49193_print.html, 17 May 2018.

[28] আহনাফ রাতুল, "শুরু হয়েছে দেশের সবচাইতে বড় আইসিটি ইভেন্ট- ডিজিটাল ওয়ার্ল্ড ২০১৬", [Online]. Available:http://www.bigganprojukti.com/?p=76344, 17 May 2018.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[29] Anandabazar Patrika, "বাংলাদেশর সবেচেয় বড় প্রযুক্তি মেলা 'ডিজিটাল ওয়াল্ড-২০১৬'", [Online]. Available: https://www.anandabazar.com/bangladesh-news/bangladesh-s-biggest-technology-fair-digital-world-2016-kick-off-bng-dgtl-1.498224#pop%20up, 17 May 2018.