An extractive text summarization technique for Bengali document(s) using K-means clustering algorithm

Conference Paper · January 2017

DOI: 10.1109/ICIVPR.2017.7890883

CITATIONS
15

READS
387

6 authors, including:

Sumya Akter
Hajee Mohammad Danesh Science and Technology University
6 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Aysa Siddika Asa
Hajee Mohammad Danesh Science and Technology University, Dinajpur
4 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Md. Palash Uddin
Deakin University
51 PUBLICATIONS **227** CITATIONS
SEE PROFILE

Md. Delowar Hossain
Kyung Hee University
27 PUBLICATIONS **124** CITATIONS
SEE PROFILE

An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm

Sumya Akter[1], Aysa Siddika Asa[2], Md. Palash Uddin[3], Md. Delowar Hossain[4], Shikhor Kumer Roy[5], and Masud Ibn Afjal[6]
Faculty of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur-5200, Bangladesh
[1]sumya.hstu@gmail.com, [2]asha.cse12@gmail.com, [3]palash_cse@hstu.ac.bd, [4]delowar.cit@gmail.com, [5]shikhorroy.cse12@gmail.com and [6]masud@hstu.ac.bd

***Abstract*— Text summarization, a field of data mining, is very important for developing various real-life applications. Many techniques have been developed for summarizing English text(s). But, a few attempts have been made for Bengali text because of its some multifaceted structure. This paper presents a method for text summarization which extracts important sentences from a single or multiple Bengali documents. The input document(s) should be pre-processed by tokenization, stemming operation etc. Then, word score is calculated by Term- Frequency/Inverse Document Frequency (TF/IDF) and sentence score is determined by summing up its constituent words' scores with its position. Cue and skeleton words have also been considered to calculate the sentence score. For single or multiple documents, K-means clustering algorithm has been applied to produce the final summary. The experimental result shows satisfactory outputs in comparison to the existing approaches possessing linear run time complexity.**

Keywords*— *data mining; text summarization; extractive summarization; bengali document(s) summarization; TF*IDF; K-means clustering algorithm

I. INTRODUCTION

Nowadays**,** the use of Internet has caused a rapid growth of electronic data which needed to process, store, and manage. Sometimes, it is difficult to find the exact information from large amount of data or big data. Big data [1] has the potential to be mined for information and data mining is essential to find out the proper information what we need. When data are being accessed from such a huge repository of e-documents, hundreds and thousands documents are retrieved through data mining. It also finds the correlations or the patterns among dozens of fields in large relational databases [2]-[3]. Data mining's roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning [4]. Data mining is thus the process used to describe knowledge in databases which is very much useful for extracting and identifying useful information and subsequent knowledge from databases. The extracted patterns from the database are then used to build data mining models, and can be used to predict performance and behavior with high accuracy. It utilizes descriptive (e.g. summarization, clustering, sequence discovery etc.) and predictive (e.g. classification, regression, time series analysis etc.) data mining approaches in order to discover hidden information [5]. As a field of data mining, text summarization is one of the most popular research areas to extract main theme from large volume of data. It is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful information. Essentially, text summarization techniques are classified as extractive and abstractive. Extractive techniques perform text summarization by selecting sentences of documents according to some criteria. Abstractive techniques attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. Sentence scoring is the most used technique for extractive text summarization. So, extractive summarization involves assigning saliency measure to some units (e.g. sentences, paragraphs) of the documents and extracting those with highest scores to include in the summary [6]. Moreover, people want to know any information in a precise way. Thus, they don't like to read big size of document with redundant data to gather information. Thus, the technique of summarizing any text document helps to find informative sentences in order to save precious time.

A. General Procedure of Text Summarization

A general procedure for extractive methods that are usually performed in three steps is discussed below [7]:

Step 1: First step creates a representation of the document. Some preprocessing such as tokenization, stop word removal, noise removal, stemming, sentence splitting, frequency computation etc. is applied here.

Step 2: In this step, sentence scoring are performed. In general, three approaches are followed:

- Word scoring–assigning scores to the most important words
- Sentence scoring–verifying sentences features such as its position in the document, similarity to the title, etc. and
- Graph scoring–analyzing the relationship between sentences.

The general methods for calculating the score of any word are word frequency, TF/IDF, upper case, proper noun, word co-occurrence, lexical similarity, etc.

The common phenomena used for scoring any sentences are Cue-phrases (''in summary'', ''in conclusion'', ''our investigation'', ''the paper describes'' and emphasizes such

978-1-5090-6004-7/17/$31.00 ©2017 IEEE

as ‘‘the best’’, ‘‘the most important’’, ‘‘according to the study’’, ‘‘significantly’’, ‘‘important’’, ‘‘in particular’’, ‘‘hardly’’, ‘‘impossible’), sentence inclusion of numerical data, sentence length, sentence centrality, sentence resemblance to the title, etc.

Also the popular graph scoring methods are text rank, bushy path of the node, aggregate similarity etc.

Step 3: In this step, high score sentences using a specific sorting order for extracting the contents are select and then final summary is generated if it is a single document summarization. For multi document summarization, the process needs to extend. Each document produces one summary and then any clustering algorithm is applied to cluster the relevant sentences of each summary to generate the final summary.

B. General Approaches for Extractive Text Summarization

Extractive summarizers [8] find out the most relevant sentences in the document. These also remove the redundant data. Extractive summarization is easier than abstractive summarization to bring out the summary. The common methods for extractive are TF/IDF method, cluster based method, graph theoretic approach, machine learning approach, LSA (Latent Semantic Analysis) method, text summarization with neural networks, automatic text summarization based on fuzzy logic, query based extractive text summarization, concept-obtained text summarization, text summarization using regression for estimating feature weights, multilingual extractive text summarization, topic-driven summarization MMR (Maximal Marginal Relevance) and centroid-based summarization, etc.

II. RELATED WORK

The previous works on single document or multi-document summarization are trying different directions to show the best result. Till now, various generic multi-document extraction based summarization techniques are already present. Most of them are for English rather than other natural languages like Bengali. In this section, we discussed some previous works on extractive text summarization. J. Zhang [9] presented an approach for multi-document text summarization using Cue-based hub-authority. It is a graph base summarization and detecting sub-topics by sentence clustering using K-nearest neighbor (KNN). Y. Ouyang [10] presented an integrated multi-document summarization approach which is based on hierarchical representation. In this paper, query relevancy and topic specificity are used for filtering process. Also it has calculated point-wise mutual information (PMI) for identifying the sub-sumption between words and high PMI is regarded as co-related. Then, hierarchical tree is constructed to produce the summarization. X. Li, J. Zhang and M. Xing [11] proposed an automatic summarization technique for Chinese text which is 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. When applying LexRank algorithm some problems are found so automatic summarization is proposed based on maximum spanning tree and sentence features to overcome the same. P. Hu, T. S. He and H. Wang [12] proposed a multi-view sentence ranking for query biased summarization. The proposed approach first constructs two base rankers to rank all the sentences in a document-set from two independent but complementary views (i.e. query-dependent view and query-independent view), and then aggregates them into a consensus one. K. Sarkar [13] presented a summarization approach based on sentence clustering for multi-documents text in which sentences are clustered using a similarity histogram based sentence-clustering algorithm to identify multiple sub-topics from the input set of related documents and selects the representative sentences from the appropriate clusters to form the final summary. A. Kogilavani [14] presented multi-document summarization using clustering and feature specific sentence extraction. V. K. Gupta [15] proposed a query focused extractive summarization approach for English text where single document summaries are combined using sentence clustering method to generate multi document summary. For clustering, semantic and syntactic similarity between sentences are used. A. R. Deshpande [16] presented another multi-document English text summarization technique using clustering method where documents are clustered using cosine similarity. A. Agrawal and U. Gupta [17] proposed an extractive clustering based technique for single document summarization. This clustering approach summarizes English text by using K-means clustering algorithm. M. A. Uddin [18] presented a multi-document text summarization for Bengali text where TF-based technology is used for extracting the most significant contents from a set of Bengali documents. Another work proposed by M. Ibrahim [19] for Bengali document summarization which is based on sentence scoring and ranking.

III. PROPOSED TECHNIQUE

The Bengali document summarizer is a natural language processing (NLP) application of data mining which is proposed to extract the most important information of the document(s). We are using sentence clustering approach to generate summary from both single and multi-documents. The proposed technique is used for summarizing Bengali document(s). In this technique, the common preprocessing steps including noise removal, tokenization, stop word removal, stemming [20] are used. TF*IDF [21] is used for calculating each word score. Then, score of each sentence is calculated by the total sum of the words with its position [19]. If the sentence contains any Cue word or skeleton word, then the score is increased by 1[19]. Next, the document is stored in a separate file with its corresponding sentences’ scores. Again for multi documents, for each document, one by one, preprocessing, word scoring and sentence scoring operations are repeated as mentioned above and also the documents are stored in the same file that is all the documents are merged. After that, the sentences are sorted in descending order and it has been considered that the highest score as the centroid 1 and the lowest score as the centroid 2 to apply the K-means clustering algorithm [1], [22]. Then, top K sentences are extracted from each cluster and the final summary is generated. Here, K sentences can be measured as 30% of sentences of the original merged document.

A. Flow Chart of the Proposed Technique

The proposed extractive technique of summarizing Bengali document(s) is illustrated with the following flow chart representation in Fig. 1.

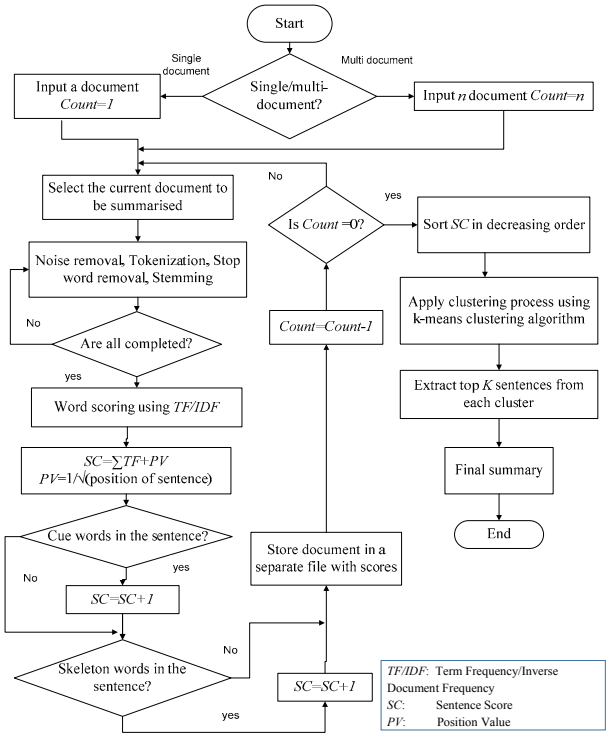

Figure 1. Extractive Bengali document(s) summarization technique

B. Pseudo-code of the Proposed Technique

TEXTSUM ( ) is the caller function that calls two procedures *Stemming ( )* & *k-means_algorithm ( )* to generate the final summary.

Procedure: TEXTSUM (*SC, COUNT, K, N, TotS, CHECK*)

i. [Start with *COUNT*]
For single document summarization,
set *COUNT*:=1.
and for multi-documents summarization,
set *COUNT:=N*, *N* is the no. of documents.

ii. [For the current document]
Count the total number of sentences, *TotS.*

iii. Set *CHECK*:=1.

iv. Repeat step (v) to (xvi) while *CHECK<= TotS.*

v. Remove noise from sentence, *S.*

vi. Tokenize each sentence *S.*

vii. Optionally remove stop word from *S.*

viii. Call procedure Stemming ( ).

ix. [Calculate the score *TF* of each word using *TF/IDF*]

$$TF = tfw_{i,s} * idfw_i$$

$$idfw_i = \log(\frac{TotS}{n_i} + 1)$$

x. [Calculate the score (*SC)* of each sentence,Sp = Position of the sentences.]

$$SC_{CHECK} = \sum TF + PV$$

$$PV = \frac{1}{\sqrt{Sp}}$$

xi. [Check *S* for cue words]
If *S* contains Cue words, then increase, *SC:=SC*+1.

xii. [Check *S* for skeleton word]
If *S* contains skeleton word, then increase, *SC:=SC*+1.

xiii. *CHECK: = CHECK*+1, goto step (iv)

xiv. Store the document in a file with scores.

xv. *COUNT:=COUNT-1*

xvi. [Check for another document to calculate sentence score]
If *COUNT!* = 0, then goto step (ii)
Else, go to step (xvii).

xvii. Sort the stored sentence scores in decreasing order.

xviii. [Cluster the document using k-means algorithm]
Call procedure k-means_algorithm ()

xix. Extract top *K* sentences from each cluster to get the final summary of the document(s).

xx. END.

Procedure: Stemming ()

i. Read token from the line.

ii. Load suffix lists from the stored file.

iii. [Check suffix list with the input token]
If the token matches with any suffix, then discard the suffix and mark token as a root word.

iv. Repeat step (i) until all token is processed.

Procedure: k-means_algorithm (*m1 & m2*, C1&C2, d1&d2, SC_m, *av1,av2*)

i. [Initialize centroid *m1 & m2*]
m1:= Highest score & *m2*:= Lowest score.

ii. [Measure distance from the centroids to each sentence *S*]
d1:= m1- SC_m, *d2:= m2-* SC_m

iii. [Check the distance either negative or not]
If *d1*<0 then *d1:= -d1*
If *d2*<0 then d2:= -d2

vi. [Create cluster]
If *d1<d2 then C1*:= SC_m
Else *C2*:= SC_m

v. [Calculating new mean]
Find average value (*av1, av2)* of clusters *C1 & C2.*

vi. [Assign the average value to the mean]
m1:= av1 & m2:= av2

vii. Repeat the steps (ii) to (vi) until the values of *m1* and *m2* in two consecutive iterations remain unchanged.

viii. Return.

1) Explanation of the Pseudo-code

The steps of the proposed method are discussed here in detail:

i. Preprocessing

The actions performed in this step are:

- Noise removal is concerned with removing header, footer, etc. from the document.
- Tokenization separates each word into lexical form. Words are separated by কমা, দাঁড়ি etc.
- The stop words are function words like এবং, অথবা, কিন্তু, অন্যথায়, কিংবা,মাত্র etc. and they may be removed.
- Stemming – A word in different forms in the same document need to be converted to their original form for simplicity like বাংলাদেশে, বাংলাদেশের, বাংলাদেশকে, বাংলাদেশেও etc. should be converted to their original form বাংলাদেশ . In the proposed technique, we used the rules for stemming any word that are illustrated in [20]. Let's consider an example: করিম কাজটি করছে. After stemming it will be করিম কাজ কর. Some examples of word stemming are shown in Table I.

TABLE I. SOME WORD STEMMING EXAMPLES

Suffix	Original words	After Stemming
ই	# এটাই, সেটাই	# এটা, সেটা
তো	# হয়তো, করলতো	# হয়, করল
কে	# এটাকে, আমাকে	# এটা, আমা
ে.ে -> া.	# হেসে, নেচে,	# হাসা, নাচা,
ে.েছিলেন -> া.	# হেসেছিলেন, নেচেছিলেন	# হাসা, নাচা

ii. Scoring Process

- Word scoring technique (TF/IDF):

This approach is used for counting the words. If there are more unique words in a given sentence, then the sentence is relatively more important [23]. The TF/IDF score is calculated as follows:

$$TF = tfw_{i,s} * idfw_i$$

$$idfw_i = \log(\frac{N}{n_i} + 1)$$

where,

TF=Term Frequency

tfw_i=Number of occurrence of the word w_i in the sentence S

$idfw_i$=Inverse document frequency

N=Total number of the sentences in the text

n_i=Number of sentences in which word w_i occurs

- Sentence scoring (SC):

$$SC_m = \sum TF + PV$$

$$PV = \frac{1}{\sqrt{Sp}}$$

Here, Sp = Position of the sentences & *PV* =Position value. For example Sp=1, if it is the first sentence of a document.

- Cue Words - If we found any cue word (e.g., মোটকথা, অবশেষে, ইতিমধ্যে, যেহেতু etc.) in any sentence, then the score of the sentence is incremented by 1.
- Skeleton word – If we found any skeleton word (e.g., headline of any document), then again the score of the sentence incremented by 1.
- Store the document in a separate file with the corresponding sentences' scores for further processing. For multi-documents, all the previous processes are applied for each document and stored in the same file where they are then merged. So for further processing, scores are sorted according to decreasing order.

iii. Applying *K*-means clustering algorithm

After sorting the scores the lowest and the highest scores are assigned as two centroids for the *K*-means algorithms and the distance from each centroid to each sentence is measured. Nearest distance from one centroid defines that cluster. Thus, two clusters are created and for next iteration, centroid values are updated. For this, the average value of each cluster is calculated and assigned them as new centroids respectively. This process is repeated until two consecutive iterations produce same result. Al last, top *K* sentences are extracted from each cluster to produce the final summary.

IV. EXPLANATION WITH AN EXAMPLE

Let's consider there are two Bengali documents to be summarized. The first document contains 10 sentences and 131 words is [24]:

বাংলাদেশের তরুণ প্রোগ্রামার ও প্রযুক্তিগত বিষয়ে ঈর্ষনীয় সাফল্যলাভকারীদের নিয়ে পৃথিবী জুড়েই প্রশংসা চলছে। বাংলাদেশ ধীরে ধীরে এগিয়ে যাচ্ছে প্রযুক্তিগত উৎকর্ষতার দিকে। এখনকার শিক্ষার্থী ও তরুণ প্রজন্ম বিজ্ঞান ও বিজ্ঞান নির্ভর পড়াশোনা নিয়ে অনেক বেশি সচেতন। সরকার তাই এই ক্ষেত্রটিকে আরো বড় একটি প্ল্যাটফর্ম হিসেবে দাঁড়া করাতে চান।বিজ্ঞান ও প্রযুক্তি নিয়ে যারা খবর রাখেন, তারা নিশ্চয়ই সামাজিক মাধ্যম কিংবা নানা ধরণের খবরে শুনেছেন ডিজিটাল ওয়ার্ল্ড ২০১৬ এর। আগামী ১৯-২১ অক্টোবর ২০১৬ বসুন্ধরা কনভেনশন সিটিতে অনুষ্ঠিত হবে দেশের সবচেয়ে বড় আইসিটি ইভেন্ট ডিজিটাল ওয়ার্ল্ড ২০১৬। এখানে থাকবে নতুন নতুন উদ্ভাবন ও প্রযুক্তিগত নানা আলোচনাও থাকছে এই আয়োজনে। এই অনুষ্ঠানে নতুন উদ্ভাবন ক্যাটাগরীতে স্কুল, কলেজ ও বিশ্ববিদ্যালয়ের বিজ্ঞানমনস্ক শিক্ষার্থীরা তাদের প্রকল্প উপস্থাপনের সুযোগ পাবে। তারা যেখানে এই প্রদর্শনীটি করবে তার নাম হচ্ছে, "ইনোভেশন জোন"। এই দায়িত্ব ও তত্ত্বাবধানে থাকছে গুগল ডেভেলপার গ্রুপস বাংলা।

The second document on the same topic contains 14 sentences and 219 words is [25]:

'ননস্টপ বাংলাদেশ' স্লোগানকে সামনে রেখে বাংলাদেশে শুরু হল তিন দিনব্যাপী তথ্য ও যোগাযোগ প্রযুক্তি বিষয়ক দেশের সবচেয়ে বড় মেলা 'ডিজিটাল ওয়ার্ল্ড-২০১৬'। বুধবার রাজধানী ঢাকার ইন্টারন্যাশনাল কনভেনশন সিটি বসুন্ধরায় (আইসিসিবি) এ মেলার উদ্বোধন করেন প্রধানমন্ত্রী শেখ হাসিনা। তিন দিনব্যাপী এই মেলা আগামী শুক্রবার পর্যন্ত প্রতি দিন সকাল ১০টা থেকে রাত ৮টা পর্যন্ত সকলের জন্য খোলা থাকবে।উদ্বোধনী অনুষ্ঠানে প্রধান অতিথির ভাষণে প্রধানমন্ত্রী শেখ হাসিনা বলেছেন, "আইসিটি ব্যবহারে তরুণ জনগোষ্ঠী নিয়ে আমরা 'লার্নিং অ্যান্ড আর্নিং' প্রকল্প চালু করেছি। প্রকল্পের আওতায় ৫০ হাজার তরুণ-তরুণীর প্রশিক্ষণের ব্যবস্থা করা হয়েছে।"ডিজিটাল সিকিউরিটি অ্যাক্ট-২০১৬ করা হচ্ছে জানিয়ে শেখ হাসিনা বলেন, "আওয়ামি লিগ সরকার দেশের স্বার্থে অর্থ ব্যয় করে সাবমেরিন কেবলের মাধ্যমে যুক্ত হয়েছে। আমরা শিক্ষা ব্যবস্থা উন্নত করতে সারা দেশে ৩০ হাজার মাল্টিমিডিয়া ক্লাস চালু করেছি। ২০১৮ সালের মধ্যে আরও দশ হাজার শেখ রাসেল ডিজিটাল ল্যাব চালু করা হবে।" ইতোমেধ্যে দেশের প্রায় সব উপ-জেলাতেই থ্রি-জি পৌঁছে গিয়েছে। আগামী ২০১৭ সালের মধ্যে ফোর-জি চালু হয়ে যাবে বলেও জানান প্রধানমন্ত্রী।বাংলাদেশ সংবাদ সংস্থা (বাসস) জানিয়েছে, ডিজিটাল বিষয়ক নয়া প্রযুক্তি ও অভিনবত্ব বিষয়ে ধারণা ও তথ্য আদানপ্রদানের জন্য এই মেলা। ৪০টি মন্ত্রণালয় ডিজিটাল বাংলাদেশ হিসেবে কী কী পরিষেবা দিচ্ছে তার খুঁটিনাটি তুলে ধরা হবে এই মেলায়। এতে শতাধিক বেসরকারি প্রতিষ্ঠান তাদের ডিজিটাল কার্যক্রম তুলে ধরবে। তিন দিনব্যাপী মেলায় মাইক্রোসফ্ট, ফেসবুক, একসেন্চার, বিশ্বব্যাঙ্ক, জেডটিই, হুয়াওয়ে-সহ বিশ্ব প্রতিষ্ঠানের ৪৩ জন বিদেশি বক্তা-সহ দুই শতাধিক বক্তা ১৮টি সেশনে অংশ নেবেন।

The score calculation of the words for the first sentence of the first document is shown in Table II.

TABLE II. WORD SCORE OF THE FIRST SENTENCE

Words	Stemming	Number of occurrence of words w_i in sentence(s) ($tfw_{i,s}$)	Number of sentence in which w_i occurs (n_i)	Score of each word, *TF*=$tfw_{i,s}$* $\log(\frac{N}{n_i}+1)$
বাংলাদেশের	বাংলাদেশ	1	1	1.04
তরুণ	তরুণ	1	2	0.78
প্রোগ্রামার	প্রোগ্রাম	1	1	1.04
ও	ও	1	6	0.43
প্রযুক্তিগত	প্রযুক্তি	1	3	0.64
বিষয়ে	বিষয়	1	1	1.04
ঈর্ষনীয়	ঈর্ষা	1	1	1.04
সাফল্যলাভকারীদের	সাফল্য	1	1	1.04
নিয়ে	নিয়ে	1	3	0.64
পৃথিবী	পৃথিবী	1	1	1.04
জুড়েই	জুড়ে	1	1	1.04
প্রশংসা	প্রশংসা	1	1	1.04
চলছে	চল	1	1	1.04

Similarly, using TF/IDF all the word scores are calculated. Then, the score of every sentence is calculated by summing up the constituent words' scores with their position using the mentioned formula. Also, the score of a sentence is increased when it contains any cue word or skeleton word or both. After getting the sentences' score of all documents, they have been sorted in a merged file which is shown in Table III for the considered example of two documents having 24 sentences in total.

TABLE III. MERGED FILE WITH SENTENCE SCORES

Sentence score notation	Score	Sentence
SC(1)	28.36	’’ ডিজিটাল সিকিউরিটি অ্যাক্ট-২০১৬ করা হচ্ছে জানিয়ে শেখ হাসিনা বলেন, ‘‘আওয়ামি লিগ সরকার দেশের স্বার্থে অর্থ ব্যয় করে সাবমেরিন কেবলের মাধ্যমে যুক্ত হয়েছে।
SC(2)	26.14	তিন দিনব্যাপী মেলায় মাইক্রোসফ্ট, ফেসবুক, একসেন্চার, বিশ্বব্যাঙ্ক, জেডটিই, হুয়াওয়ে-সহ বিশ্ব প্রতিষ্ঠানের ৪৩ জন বিদেশি বক্তা-সহ দুই শতাধিক বক্তা ১৮টি সেশনে অংশ নেবেন।
SC(3)	24.77	‘ননস্টপ বাংলাদেশ’ স্লোগানকে সামনে রেখে বাংলাদেশে শুরু হল তিন দিনব্যাপী তথ্য ও যোগাযোগ প্রযুক্তি বিষয়ক দেশের সবচেয়ে বড় মেলা ‘ডিজিটাল ওয়ার্ল্ড-২০১৬’।
SC(4)	24.68	তিন দিনব্যাপী এই মেলা আগামী শুক্রবার পর্যন্ত প্রতি দিন সকাল ১০টা থেকে রাত ৮টা পর্যন্ত সকলের জন্য খোলা থাকবে।
SC(5)	24.50	উদ্বোধনী অনুষ্ঠানে প্রধান অতিথির ভাষণে প্রধানমন্ত্রী শেখ হাসিনা বলেছেন, ‘‘আইসিটি ব্যবহারে তরুণ জনগোষ্ঠী নিয়ে আমরা ‘লার্নিং অ্যান্ড আর্নিং’ প্রকল্প চালু করেছি।
SC(6)	21.52	বাংলাদেশ সংবাদ সংস্থা (বাসস) জানিয়েছে, ডিজিটাল বিষয়ক নয়া প্রযুক্তি ও অভিনবত্ব বিষয়ে ধারণা ও তথ্য আদানপ্রদানের জন্য এই মেলা।
SC(7)	20.59	৪০টি মন্ত্রণালয় ডিজিটাল বাংলাদেশ হিসেবে কী কী পরিষেবা দিচ্ছে তার খুঁটিনাটি তুলে ধরা হবে এই মেলায়।
SC(8)	20.26	বিজ্ঞান ও প্রযুক্তি নিয়ে যারা খবর রাখেন, তারা নিশ্চয়ই সামাজিক মাধ্যম কিংবা নানা ধরণের খবরে শুনেছেন ডিজিটাল ওয়ার্ল্ড ২০১৬ এর।
SC(9)	19.39	আগামী ১৯-২১ অক্টোবর ২০১৬ বসুন্ধরা কনভেনশন সিটিতে অনুষ্ঠিত হবে দেশের সবচেয়ে বড় আইসিটি ইভেন্ট – ডিজিটাল ওয়ার্ল্ড ২০১৬।
SC(10)	17.75	বুধবার রাজধানী ঢাকার ইন্টারন্যাশনাল কনভেনশন সিটি বসুন্ধরায় (আইসিসিবি) এ মেলার উদ্বোধন করেন প্রধানমন্ত্রী শেখ হাসিনা।
SC(11)	15.67	আমরা শিক্ষা ব্যবস্থা উন্নত করতে সারা দেশে ৩০ হাজার মাল্টিমিডিয়া ক্লাস চালু করেছি।
SC(12)	15.37	২০১৮ সালের মধ্যে আরও দশ হাজার শেখ রাসেল ডিজিটাল ল্যাব চালু করা হবে।
SC(13)	14.87	এখনকার শিক্ষার্থী ও তরুণ প্রজন্ম বিজ্ঞান ও বিজ্ঞান নির্ভর পড়াশোনা নিয়ে অনেক বেশি সচেতন।
SC(14)	14.55	এই অনুষ্ঠানে নতুন উদ্ভাবন ক্যাটাগরীতে স্কুল, কলেজ ও বিশ্ববিদ্যালয়ের বিজ্ঞানমনস্ক শিক্ষার্থীরা তাদের প্রকল্প উপস্থাপনের সুযোগ পাবে।
SC(15)	13.25	আগামী ২০১৭ সালের মধ্যে ফোর-জি চালু হয়ে যাবে বলেও জানান প্রধানমন্ত্রী।
SC(16)	12.85	বাংলাদেশের তরুণ প্রোগ্রামার ও প্রযুক্তিগত বিষয়ে ঈর্ষনীয় সাফল্যলাভকারীদের নিয়ে পৃথিবী জুড়েই প্রশংসা চলছে।
SC(17)	11.67	সরকার তাই এই ক্ষেত্রটিকে আরো বড় একটি প্ল্যাটফর্ম হিসেবে দাঁড়া করাতে চান।
SC(18)	11.53	এখানে থাকবে নতুন নতুন উদ্ভাবন ও প্রযুক্তিগত নানা আলোচনাও থাকছে এই আয়োজনে।
SC(19)	11.03	প্রকল্পের আওতায় ৫০ হাজার তরুণ-তরুণীর প্রশিক্ষণের ব্যবস্থা করা হয়েছে।
SC(20)	10.45	বাংলাদেশ ধীরে ধীরে এগিয়ে যাচ্ছে প্রযুক্তিগত উৎকর্ষতার দিকে।
SC(21)	9.74	’’ ইতোমেধ্যে দেশের প্রায় সব উপ-জেলাতেই থ্রি-জি পৌঁছে গিয়েছে।
SC(22)	9.68	তারা যেখানে এই প্রদর্শনীটি করবে তার নাম হচ্ছে, ‘‘ইনোভেশন জোন’’।
SC(23)	9.47	এতে শতাধিক বেসরকারি প্রতিষ্ঠান তাদের ডিজিটাল কার্যক্রম তুলে ধরবে।
SC(24)	8.25	এই দায়িত্ব ও তত্ত্বাবধানে থাকছে গুগল ডেভেলপার গ্রুপস বাংলা।

Then, K-means clustering algorithm has been applied with *m1*=28.36 and *m2*=8.25 as the centroids. As discussed above, the final iteration is shown in Table IV.

TABLE IV. FINAL ITERATION

Centroid	Cluster	Score	Sentence
m1 = 23.25	Cluster-1	29.03	’’ ডিজিটাল সিকিউরিটি অ্যাক্ট-২০১৬ করা হচ্ছে জানিয়ে শেখ হাসিনা বলেন, ‘‘আওয়ামি লিগ সরকার দেশের স্বার্থে অর্থ ব্যয় করে সাবমেরিন কেবলের মাধ্যমে যুক্ত হয়েছে।
		26.78	তিন দিনব্যাপী মেলায় মাইক্রোসফ্ট, ফেসবুক, একসেন্চার, বিশ্বব্যাঙ্ক, জেডটিই, হুয়াওয়ে-সহ বিশ্ব প্রতিষ্ঠানের ৪৩ জন বিদেশি বক্তা-সহ দুই শতাধিক বক্তা ১৮টি সেশনে অংশ নেবেন।
		25.35	‘ননস্টপ বাংলাদেশ’ স্লোগানকে সামনে রেখে বাংলাদেশে শুরু হল তিন দিনব্যাপী তথ্য ও যোগাযোগ প্রযুক্তি বিষয়ক দেশের সবচেয়ে বড় মেলা ‘ডিজিটাল ওয়ার্ল্ড-২০১৬’।
		25.26	তিন দিনব্যাপী এই মেলা আগামী শুক্রবার পর্যন্ত প্রতি দিন সকাল ১০টা থেকে রাত ৮টা পর্যন্ত সকলের জন্য খোলা থাকবে।
		25.09	উদ্বোধনী অনুষ্ঠানে প্রধান অতিথির ভাষণে প্রধানমন্ত্রী শেখ হাসিনা বলেছেন, ‘‘আইসিটি ব্যবহারে তরুণ জনগোষ্ঠী নিয়ে আমরা ‘লার্নিং অ্যান্ড আর্নিং’ প্রকল্প চালু করেছি।
		22.07	বাংলাদেশ সংবাদ সংস্থা (বাসস) জানিয়েছে, ডিজিটাল বিষয়ক নয়া প্রযুক্তি ও অভিনবত্ব বিষয়ে ধারণা ও তথ্য আদানপ্রদানের জন্য এই মেলা।
		21.08	৪০টি মন্ত্রণালয় ডিজিটাল বাংলাদেশ হিসেবে কী কী পরিষেবা দিচ্ছে তার খুঁটিনাটি তুলে ধরা হবে এই মেলায়।
		20.26	বিজ্ঞান ও প্রযুক্তি নিয়ে যারা খবর রাখেন, তারা নিশ্চয়ই সামাজিক মাধ্যম কিংবা নানা ধরণের খবরে শুনেছেন ডিজিটাল ওয়ার্ল্ড ২০১৬ এর।
		19.39	আগামী ১৯-২১ অক্টোবর ২০১৬ বসুন্ধরা কনভেনশন সিটিতে অনুষ্ঠিত হবে দেশের সবচেয়ে বড় আইসিটি ইভেন্ট – ডিজিটাল ওয়ার্ল্ড ২০১৬।
		18.16	বুধবার রাজধানী ঢাকার ইন্টারন্যাশনাল কনভেনশন সিটি বসুন্ধরায় (আইসিসিবি) এ মেলার উদ্বোধন করেন প্রধানমন্ত্রী শেখ হাসিনা।
m2 = 11.36	Cluster-2	16.03	আমরা শিক্ষা ব্যবস্থা উন্নত করতে সারা দেশে ৩০ হাজার মাল্টিমিডিয়া ক্লাস চালু করেছি।
		15.73	২০১৮ সালের মধ্যে আরও দশ হাজার শেখ রাসেল ডিজিটাল ল্যাব চালু করা হবে।
		14.87	এখনকার শিক্ষার্থী ও তরুণ প্রজন্ম বিজ্ঞান ও বিজ্ঞান নির্ভর পড়াশোনা নিয়ে অনেক বেশি সচেতন।
		14.55	এই অনুষ্ঠানে নতুন উদ্ভাবন ক্যাটাগরীতে স্কুল, কলেজ ও বিশ্ববিদ্যালয়ের বিজ্ঞানমনস্ক শিক্ষার্থীরা তাদের প্রকল্প উপস্থাপনের সুযোগ পাবে।
		13.56	আগামী ২০১৭ সালের মধ্যে ফোর-জি চালু হয়ে যাবে বলেও জানান প্রধানমন্ত্রী।
		12.85	বাংলাদেশের তরুণ প্রোগ্রামার ও প্রযুক্তিগত বিষয়ে ঈর্ষনীয় সাফল্যলাভকারীদের নিয়ে পৃথিবী জুড়েই প্রশংসা চলছে।
		11.67	সরকার তাই এই ক্ষেত্রটিকে আরো বড় একটি প্ল্যাটফর্ম হিসেবে দাঁড়া করাতে চান।
		11.53	এখানে থাকবে নতুন নতুন উদ্ভাবন ও প্রযুক্তিগত নানা আলোচনাও থাকছে এই আয়োজনে।
		11.28	প্রকল্পের আওতায় ৫০ হাজার তরুণ-তরুণীর প্রশিক্ষণের ব্যবস্থা করা হয়েছে।
		10.45	বাংলাদেশ ধীরে ধীরে এগিয়ে যাচ্ছে প্রযুক্তিগত উৎকর্ষতার দিকে।
		9.97	’’ ইতোমেধ্যে দেশের প্রায় সব উপ-জেলাতেই থ্রি-জি পৌঁছে গিয়েছে।
		9.71	এতে শতাধিক বেসরকারি প্রতিষ্ঠান তাদের ডিজিটাল কার্যক্রম তুলে ধরবে।
		9.69	তারা যেখানে এই প্রদর্শনীটি করবে তার নাম হচ্ছে, ‘‘ইনোভেশন জোন’’।
		8.25	এই দায়িত্ব ও তত্ত্বাবধানে থাকছে গুগল ডেভেলপার গ্রুপস বাংলা।

After extracting top 5 (*K*) sentences from each cluster, the finally produced summary, which contains 10 sentences and 173 words, is shown below:

’’ ডিজিটাল সিকিউরিটি অ্যাক্ট-২০১৬ করা হচ্ছে জানিয়ে শেখ হাসিনা বলেন, ‘‘আওয়ামি লিগ সরকার দেশের স্বার্থে অর্থ ব্যয় করে সাবমেরিন কেবলের মাধ্যমে যুক্ত হয়েছে। তিন দিনব্যাপী মেলায় মাইক্রোসফ্ট, ফেসবুক, একসেন্চার, বিশ্বব্যাঙ্ক, জেডটিই, হুয়াওয়ে-সহ বিশ্ব প্রতিষ্ঠানের ৪৩ জন বিদেশি বক্তা-সহ দুই শতাধিক বক্তা ১৮টি সেশনে অংশ নেবেন। ‘ননস্টপ বাংলাদেশ’ স্লোগানকে সামনে রেখে বাংলাদেশে শুরু হল তিন দিনব্যাপী তথ্য ও যোগাযোগ প্রযুক্তি বিষয়ক দেশের সবচেয়ে বড় মেলা ‘ডিজিটাল ওয়ার্ল্ড-২০১৬’। তিন দিনব্যাপী এই মেলা আগামী শুক্রবার পর্যন্ত প্রতি দিন সকাল ১০টা থেকে রাত ৮টা পর্যন্ত সকলের জন্য খোলা থাকবে।উদ্বোধনী অনুষ্ঠানে প্রধান অতিথির ভাষণে প্রধানমন্ত্রী শেখ হাসিনা বলেছেন, ‘‘আইসিটি ব্যবহারে তরুণ জনগোষ্ঠী নিয়ে আমরা ‘লার্নিং অ্যান্ড আর্নিং’ প্রকল্প চালু করেছি।আমরা শিক্ষা ব্যবস্থা উন্নত করতে সারা দেশে ৩০ হাজার মাল্টিমিডিয়া ক্লাস চালু করেছি।২০১৮ সালের মধ্যে আরও দশ হাজার শেখ রাসেল ডিজিটাল ল্যাব চালু করা হবে। এখনকার শিক্ষার্থী ও তরুণ প্রজন্ম বিজ্ঞান ও বিজ্ঞান নির্ভর পড়াশোনা নিয়ে অনেক বেশি সচেতন।এই অনুষ্ঠানে নতুন উদ্ভাবন ক্যাটাগরীতে স্কুল, কলেজ ও বিশ্ববিদ্যালয়ের বিজ্ঞানমনস্ক শিক্ষার্থীরা তাদের প্রকল্প উপস্থাপনের সুযোগ পাবে। আগামী ২০১৭ সালের মধ্যে ফোর-জি চালু হয়ে যাবে বলেও জানান প্রধানমন্ত্রী।

V. EXPERIMENTAL RESULT AND DISCUSSION

The proposed extractive Bengali document(s) summarization technique is implemented using the IDE for Java application,

"Netbeans IDE 8.0", The performance analysis is performed in a 2.50 GHz Intel® core™ i5 CPU with 4GB RAM running Windows 7 ultimate operating system. The comparison of the proposed technique with the existing approaches is shown in Table V. So, the proposed technique summarizes both single and multiple Bengali documents though noise removing, tokenizing and stemming, scoring each word by TF/IDF and then each sentence for applying the K-means clustering algorithm.

TABLE V. COMPARISON WITH EXISTING METHODS

Author name/ Technique	Language type	Document type	Major Operations
K. Sarkar [13]	English	Multiple	Preprocessing, Clustering using cosine similarity, Cluster ordering
T.J. Siddiqui [15]	English	Multiple	Preprocessing, Sentence scoring (Feature based), Clustering using syntactic & semantic similarity
A. R. Deshpand [21]	English	Multiple (Query based)	Preprocessing (TF*IDF), Sentence scoring (Feature based), Clustering using cosine similarity
A. Agrawal [17]	English	Single	Word score (TF*IDF), Sentence scoring, K-means clustering
M. A. Uddin [18]	Bengali	Multiple	Preprocessing, Sentence scoring (TTF), Cosine Similarity measure, A* algorithm
M. I. A. Efat [19]	Bengali	Single	Preprocessing, Sentence scoring (Feature based), Sentence ranking
Proposed Technique	Bengali	Single or Multiple	Preprocessing (Noise removal, Tokenization, Stemming), Word scoring (TF/IDF), Sentence Scoring, K-means clustering algorithm

Several experiments have been conducted to evaluate the proposed technique. Some experiments are evaluated on single document and some are on multi-documents summarization. The proposed technique produced final summary from single or multiple documents which has 30% sentences of the original merged document. The proposed technique simply gives expected performance in comparison to the existing approaches. The time complexity of the proposed technique is $\theta(n)$, which offers linearity. The main pitfall is that sometimes, the sequence of summarized sentences is not synchronized. However, the technique can be applied in various summarizing fields like summarizing similar articles from different newspapers, blogs, books etc.

VI. CONCLUTION

In this paper, an extractive-based Bengali text summarization technique has been proposed both for single or multiple documents. In this summarization, some important sentences are extracted from the original document(s). We have compared the results with different extractive techniques and also measured the run-time complexity that shows the performance of the proposed technique is improved. According to the result of the proposed technique, we can conclude that it reduces the redundancy and provides better summarization. How to measure similarities is also a crucial issue in sentence clustering based summarization approach. The better similarity measure will improve the clustering performance and this may improve the summarization performance. We can measure the relevancy of the sentences using syntactic and semantic similarity in future.

REFERENCES

[1] Unnamed, Big Data. [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[2] R. M. Chezian, Ahilandeeswari. G., "A Survey on Approaches for Frequent Item Set Mining on Apache Hadoop", *International Journal for trends in Engineering & Technology,* Vol. 3 Issue 3, India, March 2015.

[3] A. Totewar, Data mining: Concepts and Techniques. [Online]. Available http://www.slideshare.net/akannshat/data-mining-15329899, India, 2016.

[4] R. Pradheepa, K. Pavithra, "A Survey on Overview of Data Mining", *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 5, Issue 8, India, 2016.

[5] Unnamed, Data Mining-Applications & Trends. [Online]. Available: https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[6] F. El-Ghannam, and T. El-Shishtawy, "Multi-Topic Multi-Document Summarizer", *International Journal of Computer Science & Information Technology (IJCSIT*), Vol 5 No 6, India, 2013.

[7] A. Nenkova and K. McKeown, "Automatic Summarization", *Foundations and Trends® in Information Retrieval*, Vol. 5, Nos. 2–3, p. 103–233, Boston - Delft, 2011.

[8] V. Gupta, G. S, Lehal, "A Survey of Text Summarization Extractive Techniques", *Journal of Emerging technologies in web intelligence,* Vol.2, No. 3, India, 2010.

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

[10] Y. Ouyang, W. Li, Q. Lu, "An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation", *Proceedings of the ACL-IJCNLP, p.* 113-116, China, 2009.

[11] X. Li, J. Zhang, 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 (IPTC),* China, 2011.

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

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

[14] A. Kogilavani, P. Balasubramani,"Clustering and Feature specific sentence extraction based summarization of multi-documents ", *International journal of computer science & information Technology (IJCSIT),* Vol.2, No.4, India, August 2010.

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

[16] R. Kamal, Bangla-Stemmer. [Online]. Available: https://github.com/rafikamal/Bangla-Stemmer

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

[18] 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),* Bangladesh, 2014.

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

[20] A. Mhatre, Implementation of k-means algorithm in C++. [Online]. Available: http://ankurm.com/implementation-of-k-means-algorithm-in-c/

[21] A. R. Deshpande, Lobo L. M. R. J, "Text Summarization using Clustering Technique", *International Journal of Engineering Trends and Technology (IJETT)* – Vol. 4 Issue 8*,* India, 2013.

[22] A. H. Witten, Text mining. [Online]. Available: http://www.cos.ufrj.br/~jano/LinkedDocuments/_papers/aula13/04-IHW Textmining.pdf

[23] R. Ferreira, L. de S. Cabral, R. D. Lins , G. P. e Silva, F. Freitas , G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization", *ELSIVIER International Journal of Expert systems with Applications*, Netherlands, 2013.

[24] আহনাফ রাতুল (2016), শুরু হয়েছে দেশের সবচাইতে বড় আইসিটি ইভেন্ট, [Online]. Available: http://www.bigganprojukti.com/?p=76344

[25] Anondobazar (2016), বাংলাদেশে শুরু সবচেয়ে বড় প্রযুক্তি মেলা' ডিজিটাল ওয়ার্ল্ড-২০১৬. [Online]. Available: http://www.anandabazar.com/bangladesh-news/bangladesh-s-biggest-technology-fair-digital-world-2016-kick-off-bng-dgtl-1.498224