

Approaches and Trends of Automatic Bangla Text Summarization: Challenges and Opportunities

Md. Majharul Haque, Bangladesh Bank, Bangladesh

Suraiya Pervin, University of Dhaka, Bangladesh

Anowar Hossain, Brain Station 23, Bangladesh

Zerina Begum, University of Dhaka, Bangladesh

ABSTRACT

As long as the internet user is increasing, online electronic content is growing proportionally irrespective of languages. A lot of research works on English text summarization have come to light to deal with this gigantic body of online text. Unfortunately, a few works have been accomplished for Bangla though a huge number of people are involved with this language. This article has tried to explore the trend of research work on Bangla text summarization. Fourteen approaches have been briefly expounded here by addressing the pros and cons with some scope of improvement. A comparison has also been turned based on their incorporated features and evaluation results. It is expected that this article will draw the attention of more researchers in the area of Bangla text summarization and give a crystal-clear message about the opportunities to the next generation. The integrated message about all the existing methods has been depicted here to reveal the importance of Bangla text summarization. To the best of the author's knowledge, this is the first review study in this ground.

KEYWORDS

Bangla, Electronic Content, Internet User, Online Text, Text Summarization

INTRODUCTION

The quantity of online available information increases rapidly with the development of the World Wide Web (Ai, Zheng, & Zhang, 2010) and the problem of information overload is rising proportionally. People are encumbered with the enormous body of electronic contents or texts, whereas they expect brief information within the shortest time. So the automatic text summarization is needed to process the large document efficiently and scavenging useful information from it (Ferreira & Souza, 2014). The goal of automatic text summarization is to condense the source text into a shorter version with preserving its information content and overall meaning (Kumar & Salim, 2012; Gupta & Lehal, 2010; Hovy, 2005).

DOI: 10.4018/IJTD.20201001.oa

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Two main categories of text summarization algorithms are extractive and abstractive (Mani, Klein, House, & Hirschman, 2002). Extraction techniques simply copy significant sentences but abstraction requires deep natural language processing which is yet to reach a mature stage even for the English language (Ye, Chua, Kan, & Qiu, 2007). The summarization task can also be classified as a single document and multiple documents text summarization (Nenkova & McKeown, 2012). The research was first started naively on the single document but today information is found from various sources on any single topic for which multiple documents summarization is demanded (Haque, Pervin, & Begum, 2013a).

The state-of-art-works (Kumar & Salim, 2012; Gupta & Lehal, 2010) focused on text summarization in various languages which were started with English text. The automatic English text summarization has begun around five decades ago by Luhn (1958) based on term-frequency. It was extended by Baxendale (1958) by incorporating the position of sentences and cue-phrases for sentence ranking. Edmundson (1969) included three additional features as (1) cue words, (2) title or heading words, and (3) location of sentences along with term frequency. Various research works are available in the arena of English text summarization (Haque et al., 2013a, 2013b), and it has witnessed the continuous involvement of many proficient researchers. However, to this age, a few works have been presented for Bangla text summarization (Sarkar, 2012a) where most of the features have been considered from the papers of English text. There are also significant amount of review papers for English text summarization regarding the discussion on various research works (Haque et al., 2013a, 2013b) from where people can understand in which point they should focus. In these circumstances, a review study on Bangla text summarization is in need so that researchers of this ground can focus on the specific points to improve.

The contribution of this paper is as follows:

1. Draw a survey with comparative study for the fourteen approaches of Bangla text summarization with pros and cons as well as the opportunities of improvement;
2. To the best of our knowledge, all the papers of Bangla text summarization have been included here from the beginning of the research work on this ground to now. It is expected that this survey will attract more researchers in this arena and give them a clear direction about the scope of improvement;
3. It has been tried to explore the pros and cons of each paper with explicit discussion;
4. Ultimately, an analysis has been drawn for some distinguished features (used in several existing methods) to show the performance improvement for each.

Though there are some existing review papers for the Indian languages and the English language to the best of our knowledge this is the first attempt to illustrate a survey, especially for Bangla text summarization.

The rest of the paper is organized as follows: The next section presents the motivation behind Bangla text summarization and then the challenges are pointed in brief. Later on, various approaches of Bangla text summarization along with some prospects, limitations, and scope of improvement have been described. Experimental results for each feature and a comparison of these approaches have also been depicted. Finally, the conclusion is turned at last.

MOTIVATION ON BANGLA TEXT SUMMARIZATION

Bangla is the 7th most spoken language in the world from more than 3500 languages and it is the native language for 250 million people (Chowdhury, Khalil, & Chowdhury, 2000). It is the mother language of Bangladesh and the second most spoken language in India. Today, many computerized contents such as web sites, word documents, etc. have been developed in Bangla because of the large

community of Bangla-speaking people. Moreover, there are several online Bangla newspapers and more of them are coming to the scene. So, e-contents in Bangla are dramatically increasing throughout the cyber world. In these circumstances, to cope-up with this large volume of text, automatic Bangla text summarization would be an invaluable solution.

We believe the following scenario will reveal the utmost importance of Bangla text summarization: if 1 of every 10 people reads Bangla newspaper regularly than 25 million people from 250 million (who speak in Bangla) are doing so. While reading the newspaper, the Bangla text summarization system can have a valuable impact if it makes a summary of all news by one-third of the total content. We may easily assume that one of us spend at least 30 minutes regularly for reading the newspaper. So, if there will be a summary consists of one-third of the content, it will save at least 10 minutes (one-third of 30 minutes) per day for each people. Based on the assumption, for the 25 million people (who read Bangla newspaper), the system will save in total $10 \times 25 \text{ million} = 250 \text{ million minutes}$ per day which is around 475 years. Indeed, nothing more is needed to say about the impact of automatic Bangla text summarization.

It is well known that the structure of Bangla's sentence is much different from English (Chowdhury, Khalil, & Chowdhury, 2000). So, the existing methods of English text summarization can't apply to Bangla. Therefore, an efficient Bangla text summarization technique is essential for researchers, international news agencies and individuals.

CHALLENGES IN RESEARCH WORK FOR BANGLA TEXT

Challenges in the research work in the ground of Bangla text are as follows:

- Automatic computerized services are hardly available for Bangla for facilitating research work;
- The lexical database like WordNet in English (Miller, 1995) does not exist for Bangla;
- There is no database of ontological meanings for Bangla words that can be used programmatically;
- Since there are a few research works exist for Bangla language, there are a few directions regarding any problem of this field.

Some other problems have also been discussed in (Karim, Kaykobad, & Murshed, 2013; Zaman, 2015) about the research work for Bangla. Further, scope of knowledge sharing is also limited as there are few researchers in this ground. Despite these difficulties, some approaches have been proposed for Bangla text summarization. These approaches are discussed in the next section.

APPROACHES OF BANGLA TEXT SUMMARIZATION

In this section, attempts of Bangla text summarization have been depicted with their strength and weakness. The scope of improvement has also been tried to explore as follows.

In 2004, Islam and Masum (2014) presented 'Bhasa', a corpus oriented search engine and summarizer. It performs document indexing and information retrieval based on keywords using vector space retrieval model ("Vector space retrieval model", 2016) for Unicode Bangla text. Corpus files can be ranked and documents can be summarized by this method based on frequent appearance of query terms. The document is treated here as one vector and query terms are treated as different vectors to get the similarity between them. A tokenizer has been used here that can determine different terms, abbreviations, tags, sentence' boundary, headings, and titles. This method has the following modules: 1) TF-IDF (term frequency-inverse document frequency) calculation module, 2) keyword search module, and 3) summary generation module. It has utilized the concept of useful, unimportant and important words' list while ranking sentences.

Discussion: This (Islam & Masum, 2014) is the first approach for Bangla text summarization along with a search engine based on our observation. It has addressed the problem of dangling pronoun and attempted to solve in extracted summary sentences but the solution has been claimed here without giving any explanation. Even, it is not specified which modules/sub-modules of this method is for text summarization or search engine. As per the TF-IDF calculation and similarity measurement of each sentence with a given query, it is exposed that the method is effective as a search engine but not for summarization. Finally, no evaluation has been given to show the application of this method in real life.

A few years later, some techniques from the investigation of English text summarization systems were applied to summarize Bangla text by Uddin and Khan (2007). They have proposed a method by incorporating some existing methods of English as follows: 1) location method, 2) cue method, 3) title method, 4) term frequency, and 5) numerical data. They have taken 40% higher-ranked sentences from the input document as a summary. It has been found that 40% of the extract by this system has got the point 8.4 from a human professional in the range of 0 to 10.

Discussion: The remarkable point of this paper (Uddin & Khan, 2007) is to show that some features of English text summarization can also be applicable for Bangla. However, this method didn't specify the exact contribution of each feature for sentence ranking. Moreover, numerical data has been considered for sentence scoring but numerical data can be presented in words instead of digits which can be considered for improvement. While evaluating this method, the score for each system generated summary was calculated but the comparison with human-generated summary/ any model summary has not been shown.

Extraction based Bangla text summarization was again presented by Sarkar (2012a). This is an easy-to-implement approach like the method of Edmandson (1969) with the three major steps: (1) preprocessing, (2) sentence ranking, and (3) summary generation. The impact of the thematic term has been investigated and features like word-frequency, length, and position of the sentence have been utilized for sentence ranking. It was claimed that the system performs better than the LEAD baseline method (the first n words of an input article are considered as the summary in LEAD baseline method). Average unigram based recall score was found as 0.4122.

Discussion: This method (Sarkar, 2012a) is fully based on almost four decades of old English text summarization method (Edmundson, 1969) which can be upgraded by incorporating modern natural language processing techniques like sentence clustering, redundancy removal, etc. Moreover, in the evaluation, only one model summary has been used for each of the test document but more model summaries can be developed for sophisticated evaluation results (Haque, Pervin, & Begum, 2016).

In 2012, Sarkar (2012b) proposed another method by tuning each feature of his previous method (Sarkar, 2012a) for better summarization performance. This approach has four major steps (1) preprocessing (2) extraction of candidate summary sentences (3) ranking the candidate summary sentences (4) summary generation. This is also based on word-frequency, sentence position and sentence length that is similar to (Sarkar, 2012a). In this approach, some threshold points have been adjusted for the position of sentences, TF*IDF values and the minimum length of sentences. The impact of each feature has been specified with experiments for sentence ranking.

Discussion: This method (Sarkar, 2012b) has surpassed the LEAD baseline method, baseline that uses term-frequency with sentence location and the method described in (Sarkar, 2012a). All the features have been tuned here for better performance. However, this method is also based on an old English text summarization procedure (Edmundson, 1969). Moreover, the evaluation has been turned here against only one model summary where more than one model summary can be used for getting a sophisticated evaluation result (Haque, Pervin, & Begum, 2016). This system can also be upgraded by incorporating modern natural language processing techniques as discussed for the previous method.

In 2013, Efat, Ibrahim, and Kayesh (2013) introduced a method for Bangla text summarization by sentence scoring and ranking. Their system is alienated into three segments: (1) pre-processing the test document, (2) sentence scoring, and (3) generating a summary. Sentence scoring is depended on

term frequency, position, cue words and skeleton of the document. Skeleton of the document consists of the words in title and headers.

Discussion: It was stated in their paper that the system performs well when the document completely depends on a particular theme (Efati, Ibrahim, & Kayesh, 2013). So, the system can be more user-friendly by eliminating this dependency. The average accuracy of this proposed method has been claimed 83.57% against human-generated summarization which is really a good sign but they didn't give a comparison with any existing method. An experiment has also been turned to measure the contribution of each feature while sentence ranking. Nevertheless, the evaluation result is for a particular theme only which may not be comparable with other generic text summarization methods.

Abstraction based Bangla text summarization system was proposed for the first time in 2014 by Kallimani, Srinivasa, and Reddy (2014). They focused on a unified model with attribute-based Information Extraction rules and class-based templates. They have claimed for adaptation of this system over four Indian languages as Kannada, Hindi, Bangla, and Telugu. The document to be summarized is subjected to preprocessing, namely – Parts of Speech (POS) tagging and Named Entity Recognition (NER). TF/IDF rule-based classifier has also been used to categorize the document which determines the applicable classes. In this system (Kallimani, Srinivasa, & Reddy, 2014), classes are blueprints where the identified attributes are set according to this blueprint. Attributes are primary pieces of information as follows - NAME, PLACE, DOB (date of birth), DOD (date of demise), and AWARDS. The most significant part of this system is the template-based sentence generation where templates are generic structures of sentences with some gap of crucial pieces of information. The extracted attributes are mapped with the templates to generate summary sentences.

Discussion: It is well known that abstraction based English text summarization is yet in an immature stage (Ye, Chua, Kan & Qiu, 2007) though the research work on English text summarization was begun in 1958 (Luhn, 1958). In this situation, this method (Kallimani et al., 2014) has reported for Bangla abstractive summarization. Attribute extraction of this method is noticeable which is required for informative sentence generation. Nevertheless, the utilized template is creating the same structure of sentences always which can be monotonous. It is also questionable that using the template is enough or not for all types of sentences while abstraction. So, there is a scope of improvement for generating refined sentences with the identified attributes. According to their evaluation, the system achieved an average 86.24% precision, 78.93% recall, and 81.50% F-measure is an intrinsic evaluation. The evaluation that has been turned here it seemed to be for attribute extraction only. It is because the Precision and Recall value can be measured by matching for the important items extracted by the system against the important items that exist in the text.

Research work has been accomplished for multiple document text summarization for Bangla language in 2014 by Uddin, Sultana, and Alam (2014) for the first time. In this paper, a primary summary is generated at first by sentence scoring on the basis of term-frequency. It has been reported that the words are replaced with their common synonym before term frequency calculation so that different words with the same meaning will be treated as the same word. Cosine similarity for each sentence to every other sentence of primary summary has been calculated to get the relevance between them. A graph-based model is then applied with the A* (Aker, Cohn, & Gaizauskas, 2010) searching algorithm on the primary summary for creating the final gist. It has been claimed that the selection of the starting point of a summary is effective by this method. The performance evaluation has been completed against the human-generated summary. Unigram based Recall Score was found 56% and the similarity between manual and system generated summary was shown 86.60%. The relevance among three human judges has also been shown evidently in their paper to expose that a single sentence is not taken as significant or worthless by all judges equally.

Discussion: It is noticeable that this is the first task for Bangla multiple documents text summarization. This method has selected the most relevant sentence as the starting point of summary but no theoretical or practical reason has been stated behind this. Even the source of getting synonyms for each word before term-frequency calculation has not been mentioned. After selecting the final

summary sentences, there is no direction for the ordering of the sentences of different sources which is very much necessary to make the text lucid and understandable.

Apart from the previous approaches, keyphrase based summarization method outperforms for both Bangla and English text, which was proposed by (Sarkar, 2014). Keyphrases are extracted as a sequence of words from any sentence that contains no punctuation mark and stop words. All the keyphrases are ranked as per their frequency and the sentences are ranked based on position and term frequency. Summary sentences are selected in two phases. In phase-1, candidate summary sentences are chosen that contain top-ranked keyphrases. From the chosen sentences, top-ranked sentences are selected that have the position not more than fifth in place in the document. If phase-1 fails to generate a summary of user-desired length, phase-2 is activated and select more summary sentences based on the sentences' score from the rest of the sentences.

Discussion: This (Sarkar, 2014) is the first task on keyphrase-based sentence extraction for Bangla text summarization. It was claimed that keyphrases can reflect the concept of a document more clearly than words. This method has set the upper limit in the length of keyphrases but no lower limit has been set. So, an experiment can be done here to set the lower limit in the length of keyphrases for better performance. In the evaluation, this method outperforms all the existing methods of Bangla text summarization. However, the same type of method has already been introduced for English (Sarkar, 2013). For sentence scoring, only position and term frequency have been considered that have already been introduced around four decades ago (Edmundson, 1969). Today, it can be seen that a lot of significant features have been invented by various researchers for text summarization (Haque, Pervin, & Begum, 2013a, 2013b). So, the performance of this research work can be enhanced by adding more features for sentence scoring.

Sentence clustering-based Bangla news document summarization was published for the first time in 2015 (Haque, Pervin, & Begum, 2015). They have introduced sentence frequency along with term frequency. Sentences are ranked here by doing an algebraic sum of the scores of term frequency, sentence frequency, and numerical figure. Initially, sentence frequency is set to zero (0) for each and then every sentence is matched with others. If any sentence is found containing 60% terms of any other sentence, the smaller sentence is removed and the frequency of larger sentences is increased between them. Sentences are clustered according to their cosine similarity ratio and one-third top-ranked sentences are selected from each cluster. It was claimed that clustering helps for better coverage of information in summary.

Discussion: This method (Haque, Pervin, & Begum, 2015) has introduced sentence clustering for the first time in Bangla text summarization. Sentence frequency is another contribution of this research work which assists in redundancy elimination and sentence ranking. However, clustering by cosine similarity is very much conventional which works on directly matching of terms between two sentences (Yang, Cai, Zhang, & Shi, 2014). This clustering strategy can be updated by utilizing background knowledge from Banglapedia so that two different terms can be matched semantically and lexically. In the updated way, two sentences can be in the same cluster though they may have low cosine similarity. Again, the numerical figure has been counted here for sentence ranking but no strategy has been proposed to identify numerical figures if they are in words' form other than digits. Moreover, the weight of each sentence ranking feature can be set experimentally for better summarization performance. In the performance measurement against human-generated summary, the F-measure score has been found 0.632 where only 20 documents have been considered.

A well-established keyphrase based method (Sarkar, 2014) has been enhanced by Haque et al. (Haque, Pervin, & Begum, 2016) for Bangla news documents summarization. Here, the existing method (Sarkar, 2014) is scrutinized and the way of betterment is mentioned. The enhancements incorporate: (i) modifying the keyphrases selection process, (ii) including the first sentence in summary if it contains any title word and (iii) counting numerical figure which is presented in digits and words for sentence scoring. The evaluation has been drawn by considering 200 documents with 3 summaries for each (in total $3 \times 200 = 600$ summaries) using ROUGE (Recall Oriented Understudy for Gisting

Evaluation) (Lin & Hovy, 2003; “ROUGE 2.0”, 2016) score. In the evaluation, the F-measure score has been enhanced from 0.5496 to 0.6166 (ROUGE-1) and from 0.5050 to 0.5830 (ROUGE-2).

Discussion: The method (Haque, Pervin, & Begum, 2016) is an enhancement of an existing keyphrase based method presented in (Sarkar, 2014). It is remarkable that the new method significantly outperforms the existing methods. Point to be mentioned that this research work has utilized the ROUGE package (“ROUGE 2.0”, 2016) for the first time for evaluating the Bangla text summarization system. It has counted numerical figures from words for sentence scoring which was not noticed by any other methods based on our study. Nevertheless, the text form of the numerical figure with three digits contains more than one word in the Bangla language. For example, “123 – একত্ৰিশ” is written as “one hundred twenty-three – একত্ৰিশ”. In this situation, one numerical figure can be counted for two times (one hundred – একত্ৰিশ and ত্ৰিশ- twenty-three). No mechanism has been included here to handle this issue. Again, background knowledge of keyphrases (from Banglapedia) can be considered while sentence ranking for upgrading performance. Another significant point that has not been covered is dangling pronoun resolution. If a sentence is extracted where a pronoun is available but the sentence which is containing the noun of that pronoun is not included in the summary, the summary will be ambiguous.

Later on, another method has been presented by Haque, Pervin, and Begum (2017a) where pronoun replacement is accomplished for the first time to minimize the dangling pronoun from summary. After replacing pronoun by the corresponding noun, sentences are ranked by considering (i) term frequency, (ii) sentence frequency, (iii) numerical figures and (iv) title words. Dependency parsing has been introduced here for general and special tagging of unknown words based on the tag of known words. The first sentence is included in summary always if it contains any title word. It has been found from the ROUGE evaluation results that the method outperforms the four latest existing methods (Sarkar, 2012a, 2012b, 2014; Efati et al., 2013).

Discussion: In this method (Haque, Pervin, & Begum, 2017a), pronoun replacement by the corresponding noun has been utilized for the first time. The numerical figures have been considered here for both digits and words form. Their system is a rule-based system that utilized a hidden Markov model and Markov chain model. It has been claimed that 3,000 Bangla news documents have been analyzed to explore the rules. Along with the parts-of-speech tagging, they have introduced special tagging including Acronyms, Repeated words, Occupation, Name of humans and places, etc. Dependency parsing is another notable feature for boosting the tagging procedure. However, most of the rules, they have used here for dependency parsing, pronoun replacing and special tagging, have no grammatical reference which is the principal concern of this paper. Though they have identified the full human name and recalled the full name from the part of the name, there can be a chance of a high false-positive rate to treat any word accurately as the part of the name. So, it can be stated that there is a significant scope of improvement for this method.

A heuristic approach of Bengali text summarization has been proposed by Abujar, Hasan, Shahin, and Hossain (2017). They have claimed for deriving some rules of Bangla text analysis. Three phases have been accomplished here as (i) preprocessing with linguistic analysis, (ii) Prime sentence (the main leading sentence) identification by words and sentence analysis, and (iii) final processing for the betterment of summary generation. The sentence analogy matrix has been utilized and sentence imitation has been considered to omit redundant sentences. They have calculated the effective rate of words by considering the repeated distance. The first and last positional sentences of each paragraph have been treated as significant for sentence scoring. They have claimed that throughout the proposed rules and models, final processing features can generate a better quality summary from Bangla text. The evaluation has been done with the human-generated summary with only three different texts where the performance is showing almost similar to humans without mentioning the actual performance in the numerical figure.

Discussion: In the paper (Abujar et al., 2017), the relations between words and sentences have been revealed and the prime sentence has been selected with some other steps to develop a better

summarization system. Words' effect rate identification is deemed to be significant but no justification has been provided for the range of effective rate. Cue words have been considered though positive or negative cue words could be differentiated for the improvement. Imitating of sentences is taken care of so that redundancy can be minimized but a similar feature has already been proposed in (Haque, Pervin, & Begum, 2015) as sentence frequency which has not been mentioned. The identification of prime sentences can have a false-positive rate which deserves a statistical analysis. Furthermore, the method can be tested with a publicly available dataset ("Dataset", 2016) rather than human-generated summaries only. However, it has been explicitly mentioned in the paper that there are lots of improvements required for an enhanced summarizer.

Ghosh, Shahariar, and Khan (2018) proposed a Rule-based extractive summarization system by utilizing 12 features in 2018. According to their statement, major contributions are: (i) applying graph-based sentence scoring features, (ii) introducing some features for the first time as like aggregate similarity, bushy path, keyword in the sentence, presence of inverted comma and special symbol, (iii) removing redundant information from the summary. The first sentence, based on position, has been emphasized for summary generation and the importance is downgraded to the second, third and so on. The Cue-words and title words have also been taken into account for important sentence identification.

Discussion: In this method (Ghosh, Shahariar, & Khan, 2018), they have stated that 12 features have been utilized. It is appreciated that they have brought some new features for Bangla text summarization and outperformed all the existing methods in evaluation. The evaluation has been turned with a published dataset (Haque, Pervin, & Begum, 2015) with ROUGE evaluation tools. The comparison has been shown with the 5 latest existing methods. However, all the features have been equally considered without depicting any analytical result for each feature individually. Haque et al. (Haque, Pervin, & Begum, 2017a) showed that the weight of every feature should not be the same. Moreover, there is no partial implementation of the method so that the contribution of each feature can be individually distinguished. They have considered the numerical figure which is presented in digits only and ignored the figure that can be presented in words. Furthermore, the significant issue of dangling pronoun (Haque, Pervin, & Begum, 2017b) has not been addressed in their research work.

Sikder, Hossain, and Robi (2019) presented a method of Bangla text summarization by combining some mathematical and Bangla grammatical rules in 2019. It has been claimed that they have introduced the first idea of extraction method including grammatical view which is a path of abstraction. According to the paper, the main contribution of this research work is sentence relevancy, meaning analysis, joining and eliminating odd sentences. After preprocessing, sentence ranking has been done by considering Term frequency, sentence position, sentence' similarity and then 70% top-ranked sentences are selected as the primary summary. From the primary summary, sentence joining has been done by considering some Bangla grammatical rules where two or more sentences are transformed into a single sentence. While joining, the related nearest sentences are identified for each sentence of the primary summary and also distinguish the structure of all the related sentences. Finally, simplified sentences are generated for each sentence and place in the appropriate position. The evaluation of this method has been accomplished with a human-generated summary for six different documents.

Discussion: It is appreciated that the method (Sikder, Hossain, & Robi, 2019) has considered Bangla grammatical rules along with mathematical rules and introduced the path of abstraction based summarization. This has defined the ways of constructing new sentences from related consecutive sentences. This paper claimed for the first step in Bangla text abstraction whereas the abstraction based Bangla text summarization has been presented in 2014 (Kallimani et al., 2014). Though the method has been proposed for Bangla text summarization, they have claimed that their method can be extended easily for other languages also. However, it is deemed to be impossible for the claimed extension for other languages because there is a clear explanation that there are significant differences between Bangla and English based on grammatical rules (Haque, Pervin, & Begum, 2017a). Moreover, no analytical justification has been provided for the sentence positional score that the first sentence

will have top importance and downgraded for the next sentences gradually. Sentence joining is an appreciated step but an analytical review is needed to show the impact of this step. After all, the evaluation could be done by utilizing more documents instead of six documents only.

Some noticeable points about Bangla text summarization methodologies are as follows:

- Most of the utilized features have been taken from existing English text summarization;
- There is no common dataset publicly available for evaluating Bangla text summarization system. In this regard, we have generated and uploaded dataset in (“Dataset”, 2016) which can be used in any upcoming methods. This dataset has already been used for by some research works (“Dataset”, 2016). We hope that it will be helpful for researchers to evaluate their methods in future;
- Semantic knowledge-base can be implemented to help in creating proficient Bangla text summarization system;
- A lexical dictionary like WordNet (Miller, 1995) in English can be developed for Bangla text.

EXPERMENT WITH DIFFERENT FEATURES

Experiment has been done by generating summary of Bangla text document where only one input is taken which is a Bangla news document. Through the experiment, impact analysis of different features has been accomplished. In this analysis, the generated summary has been compared with three model summaries of 200 news documents each and the results of evaluation is the average results of the comparisons. Precision, Recall and F-measure are brought into play here as these have long used as important evaluation matrices in information retrieval field. If ‘A’ indicates the number of sentences retrieved by summarizer and ‘B’ indicates the number of sentences that are relevant as compared to target set, Precision, Recall and F-measure are computed based on the following equations:

$$Precision (P) = \frac{A \cap B}{A} \quad (1)$$

$$Recall (R) = \frac{A \cap B}{B} \quad (2)$$

$$F\text{-measure} = \frac{2 \times P \times R}{P + R} \quad (3)$$

To show the impact of features in the betterment of summary generation, some distinguished features have been selected to apply as follows:

1. Pronoun replacing by the corresponding noun to minimize the number of dangling pronouns;
2. Sentence ranking by:
 - a. Term frequency inverse document frequency calculation;
 - b. Sentence frequency measurement with redundancy elimination;
 - c. Counting the existence of numerical data from digits;
 - d. Counting the existence of numerical data from words;
 - e. Computing title word score;
3. Considering the first sentence especially if it contains any title word;
4. Adjustment of the coefficients of all the attributes listed in the above point (ii).

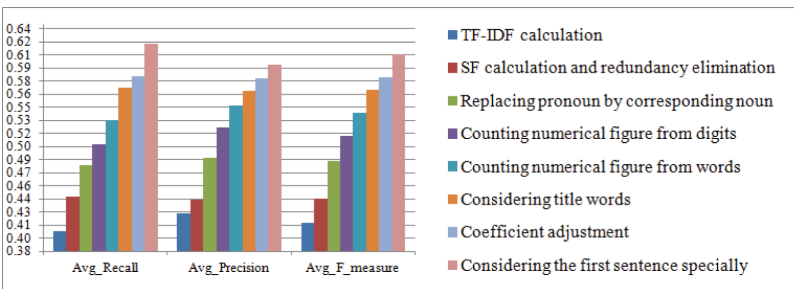
The development code has been written in PHP web based programming language where the 363 Bangla stop words have been used from the Indian Statistical Institute website (“Stop Words”, 2016). The simulation has been done in a laptop with Processor: Intel(R) Core(TM) i5-3210M CPU @2.50GHz 2.50GHz, Installed memory (RAM): 8.00 GB (7.60 GB usable), System type: Windows 7 Professional 64-bit Operating System.

The impact has been estimated as Precision, Recall, and F-measure using Equations (1), (2), and (3) respectively using a publicly available dataset (“Dataset”, 2016) of 200 news document with three summaries for each of the document (total 600 summaries). Every time, the system generated summary is compared with three model summaries of each document, and compute the average value of Precision, Recall and F-measure with ROUGE (“ROUGE 2.0”, 2016) automatic evaluation package. Point to be mentioned that the features have been implemented as like the feature mentioned in some paper as follows:

1. **Sentence Frequency Calculation:** If there are two or more sentences are found with 60% similarity, the long sentence is kept and the other is deleted as in paper (Haque, Pervin, & Begum, 2015);
2. **Replacing Pronoun by Corresponding Noun:** The pronoun has been replaced by the corresponding noun so that the related noun and pronoun will be treated as same word. Moreover, it will impact in the word frequency calculation. This has been implemented with the rule based replacement of pronoun as in (Haque, Pervin, & Begum, 2017b);
3. **Count Numerical Figure From Digits:** The numerical figure can be presented in digits which has been counted with the pattern recognition addressed in (Haque, Pervin, & Begum, 2016);
4. **Count Numerical Figure From Words and Digits:** The numerical figure can be presented in both words and digits which has been counted with the pattern recognition addressed in (Haque, Pervin, & Begum, 2016);
5. **Considering Title Words:** The title words score has been used in several methods for sentence ranking as in (Sarkar, 2012a; Haque et al., 2013a, 2013b; Islam & Masum, 2014; Efat, Ibrahim, & Kayesh, 2013);
6. **Coefficients Adjustment:** Sentence ranking has been done in several Bangla text summarization methods using some parameters like numerical figure, sentence frequency, title word, term frequency, etc. For these parameters, the impact of all these are not same according to the paper (Sarkar, 2012b, Haque, Pervin, & Begum, 2015) for which the coefficients of each parameters are adjusted for getting better performance;
7. **Considering the First Sentence Specially:** The first sentence of every document is considered as important (Haque, Pervin, & Begum, 2015) which is also considered by us.

In Figure 1, all the above features has been added one by one and the utilized features in each step include all the features of the previous step(s).

Figure 1. Step by step improvement of performance for including each feature



After generating summary by using only term-frequency, the F-measure score has been found as 0.4124 using the same dataset (“Dataset”, 2016). After incorporating each feature the F-measure score has been raised as mentioned in Table 1.

COMPARISON AMONG VARIOUS APPROACHES

Fourteen approaches have been discussed in the previous section with their pros and cons. Table 2 turns the comparison among these approaches based on their incorporated features and evaluation results.

Table 1. Percentage of performance improvement

SN#	Features	Improvement
1	Sentence frequency calculation	6.71%
2	Replacing pronoun by corresponding noun	9.66%
3	Count numerical figure from digits	6.15%
4	Count numerical figure from words and digits	5.11%
5	Considering title words	4.96%
6	Coefficients adjustment	2.47%
7	Considering the first sentence specially	4.47%

Table 2. Comparison among fourteen approaches of Bangla text summarization

Sn#	Researcher (s), Year	Incorporated Distinguished Features	Remarks	Evaluation Result
1	Islam and Masum (2014)	i) Term frequency ii) Useful word list with Important and unimportant word list	This method has keyword search module for summary generation where keywords are selected on the basis of tf-idf and list of useful, important and unimportant words.	No evaluation has been drawn.
2	Uddin and Khan (2007)	i) Using location method ii) Using Cue method iii) Using Title method iv) Term frequency	This research work has shown that some features of English text summarization can be used for Bangla text.	Got 8.4 from human professional in the range of 0 to 10 point with 40% extraction.
3	Sarkar (2012a)	i) Term frequency ii) Sentence length iii) Sentence position	The impact of thematic term has been investigated here and some statistical measures have been incorporated for sentence scoring.	Unigram based recall score is 0.4122.
4	Sarkar (2012b)	i) Term frequency ii) Sentence length iii) Sentence position	It was claimed that the features used here in more effective way for news document summarization than in the previous method (Sarkar, 2012a). In this approach, some threshold points have been adjusted for position of sentences, TF-IDF values and minimum length for selecting summary sentences.	Precision, Recall and F-measure values have been claimed 0.3659, 0.5064 and 0.4169 respectively.

continued on following page

Table 2. Continued

Sn#	Researcher (s), Year	Incorporated Distinguished Features	Remarks	Evaluation Result
5	Efat et al. (2013)	i) Term frequency ii) Sentence position iii) Skeleton of document	Their system is alienated into three segments as pre-processing the test document, sentence scoring and summarization based on sentences' score.	The average accuracy of this proposed method has been found 83.57% against human generated summary.
6	Kallimani et al. (2014)	i) Parts-of-speech tagging ii) Named entity recognition iii) Utilizing template of sentence iv) Abstraction	Input document is categorized to apply specific classes. Some attributes are extracted from the document and mapped with the template of sentence for summary sentence generation.	The system achieved an average of 86.24% precision, 78.93% recall, and 81.5% F-measure while intrinsic evaluation.
7	Uddin et al. (2014)	i) Term frequency ii) Cosine similarity among sentences iii) A* (Aker et al., 2010) searching algorithm iv) Multi-document text summarization	This is a multi-document text summarization system. A primary summary is generated at first by sentence scoring. A graph based model is then applied with the A* (Aker et al., 2010) searching algorithm on the primary summary for creating the final gist.	Unigram based Recall Score was claimed as 56%.
8	Sarker (2014)	i) Keyphrase extraction ii) Sentence position iii) Term frequency	This is a keyphrase-based sentence extraction method for both Bangla and English document. Here, keyphrases are sequence of words without any stop words and punctuation marks. Two phases approach have been applied here. First phase will select sentence on the basis of top ranked keyphrases and sentence score. Second phase is activated if summary can't be created in first phase and select more sentences based on score.	It was claimed that this method outperforms existing methods (Sarkar, 2012a, 2012b) of Bangla text summarization. The F-measure score has been found 0.4242 in the evaluation for Bangla text summarization.
9	Haque et al. (2015)	i) Term frequency ii) Sentence frequency iii) Counting Numerical figure iv) Sentence clustering	In this method, sentences are ranked using term frequency, sentence frequency and counting numerical figure. Initially, sentence frequency is set to zero (0) for each. Then, if any sentence is found containing 60% terms of any other, smaller sentence is removed and the frequency of larger sentence is increased between them. Sentence clustering has been utilized here to carry diversified information in summary. After clustering, one third top ranked sentences are selected.	Precision, Recall and F-measure values have been claimed as 0.608, 0.664 and 0.632 respectively.
10	Haque et al. (2016)	i) Setting minimum length of keyphrases ii) Considering the first sentence iii) Counting numerical figure from both digits and words.	This method is an enhancement of an existing keyphrase based method of (Sarkar, 2014). The enhancements include: i) setting the minimum length for keyphrases, ii) considering the first sentence specially and iii) counting numerical figure from words and digits for sentence scoring.	As per ROUGE-1 score, the value of F-measure was found as 0.6166 (F-measure for the previous method (Sarkar, 2014): 0.5495) in the evaluation with same dataset mentioned in (Haque, Pervin, Begum, 2016).

continued on following page

Table 2. Continued

Sn#	Researcher (s), Year	Incorporated Distinguished Features	Remarks	Evaluation Result
11	Haque et al. (2017)	<ul style="list-style-type: none"> i) Term frequency ii) Sentence frequency iii) Counting Numerical figure both in numeric and word forms iv) Title words v) General & Special tagging vi) Dependency parsing vii) Replacement of pronouns 	In this method, two significant contributions are (i) pronoun replacement to solve the issue of dangling pronoun and (ii) dependency parsing to enhance the tagging procedure. After replacing pronoun by corresponding noun, sentences are ranked where special tagging has been utilized so that important sentences will be in higher rank. Moreover, the first sentence is included in summary always if it contains any title word.	F-measure scores have been found as 0.6003 and 0.5708 for ROUGE-1 and ROUGE-2 respectively by the evaluation with a publicly available dataset from ("Dataset", 2016). Comparison has been turned with the four existing methods (Sarkar, 2012a; 2012b; Efat et al., 2013; Sarkar, 2014) where this method has outperformed others.
12	Abujar et al. (2017)	<ul style="list-style-type: none"> i) Term frequency ii) Words and sentences analysis iii) Numerical value identification iv) Sentence position and length v) Title words vi) Cue words vii) Word effect rate viii) Prime sentence identification ix) Aggregate similarity measurement x) Detection of imitating of sentences 	In this method the authors have claimed for deriving some rules of Bangla text analysis. After preprocessing with linguistic analysis, prime sentence has been identified by words and sentence analysis. They have calculated the effect rate of words by considering repeated distance and aggregate similarity has been measured to eliminate the redundancy. Moreover, sentence position, length, cue words, numerical value, term frequency and title words have been considered for sentence ranking.	The evaluation has been done with human-generated summary with only three different texts where the performance is showing almost similar to human without mentioning the actual performance in numerical figure. No comparison has been shown against any existing methods.
13	Ghosh et al. (2018)	<ul style="list-style-type: none"> i) Aggregate similarity ii) BushiPath iii) Term frequency inverse sentence frequency iv) Keywords v) Sentence position vi) Title words vii) Cue words viii) Numerical value ix) Inverted comma x) Special symbol xi) Date format xii) URL/Email address 	This is a rule based extraction based text summarization which has utilized 12 features for sentence ranking. Major contributions of this method includes: (i) applying graph based sentence scoring features, (ii) introducing some features for the first time as like aggregate similarity, bushy path, keyword in sentence, presence of inverted comma and special symbol, (iii) removing redundant information from summary.	F-measure scores have been found as 0.6276 for ROUGE-1 evaluation result with a publicly available dataset from ("Dataset", 2016). Comparison has been turned with the five existing methods (Sarkar, 2012a; 2012b; Efat et al., 2013; Sarkar, 2014; Haque et al., 2017) where this method has outperformed others.

continued on following page

Table 2. Continued

Sn#	Researcher (s), Year	Incorporated Distinguished Features	Remarks	Evaluation Result
14	Sikder et al. (2019)	<ul style="list-style-type: none"> i) Term frequency ii) Sentence relevancy iii) Position iv) Bangla grammatical rules v) Primary summary generation vi) Sentence simplification rules vii) Sentence joining and linking viii) Abstraction 	This method has considered Bangla grammatical rules along with mathematical rules and introduced the path of abstraction based summarization. It has been mentioned that the main contribution of this research work is sentence relevancy, meaning analysis, joining and eliminating odd sentences. After sentence ranking, 70% top ranked sentences are selected as the primary summary. From these sentences, sentences are joined, redundancy eliminated and simplified to generate final summary.	The evaluation has been accomplished with human-generated summary for six documents only. The result of evaluation has been depicted in graph without mentioning the actual numerical value about the performance.

CONCLUSION

In this paper, fourteen approaches of Bangla text summarization have been described where thirteen methods are for the single document and one is for multiple documents. Two methods are there for abstraction based text summarization. The trend in the research work of Bangla text summarization has been tried to explore in brief. A comparison table has also been drawn to view the similarity and differences among them. The strength and weaknesses of all the methods have been discussed along with the scope of improvement. It has been indicated with the reference that most of the incorporated features in various existing methods of Bangla text summarization were collected from the methods of English text but with a different angle, because the structure of Bangla is different from English. After all, this little number of efforts is raising hope for a more sophisticated methodology soon. It is also expected that this review paper will help the next generation to know the basement of research works in Bangla text summarization and to get the direction of future works.

Our future work is to do an impact analysis of different features to identify their ratio of workability for different pattern of documents for Bangla text summarization.

ACKNOWLEDGMENT

This research work is funded by a Fellowship Scholarship from Information and Communication Technology Division, Government of the Peoples Republic of Bangladesh. There is also a valuable support from the Central Bank of Bangladesh.

REFERENCES

- Abujar, S., Hasan, M., Shahin, M. S. I., & Hossain, S. A. (2017). A Heuristic Approach of Text Summarization for Bengali Documentation. *8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-8. doi:10.1109/ICCCNT.2017.8204166
- Ai, D., Zheng, Y., & Zhang, D. (2010). Automatic text summarization based on latent semantic indexing. *Journal of Artificial Life and Robotics*, 15(1), 25–29. doi:10.1007/s10015-010-0759-x
- Aker, A., Cohn, T., & Gaizauskas, R. (2010). Multi-document summarization using A* search and discriminative training. *Conference on Empirical Methods in Natural language Processing*, 482-491.
- Banglapedia. (2003). *National Encyclopedia of Bangladesh*. Asiatic Society of Bangladesh.
- Baxendale, P. B. (1958). Machine-made Index for Technical Literature -An Experiment. *IBM Journal of Research and Development*, 2(4), 354–361. doi:10.1147/rd.24.0354
- Chowdhury, M., Khalil, I., & Chowdhury, M. H. (2000). *Bangla VasarByakaran*. Ideal Publication.
- Dataset for evaluating Bangla text summarization system. (n.d.). *Bangla Natural Language Processing Community*. Retrieved at Jan, 2016 from <http://bnlpc.org/research.php>
- Dhanya, P.M., & Jathavedan, M. (2013). Comparative Study of Text Summarization in Indian Languages. *International Journal of Computer Applications*, 75(6), 17-21.
- Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2), 264–285. doi:10.1145/321510.321519
- Efat, M. I. A., Ibrahim, M., & Kayesh, H. (2013). Automated Bangla Text Summarization by Sentence Scoring and Ranking. In *International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE. doi:10.1109/ICIEV.2013.6572686
- Ferreira, R., & Souza, L. D. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Journal of Expert Systems with Applications, Elsevier*, 41(13), 5780–5787. doi:10.1016/j.eswa.2014.03.023
- Ghosh, P., Shahariar, R., & Khan, M. (2018). A Rule Based Extractive Text Summarization Technique for Bangla News Documents. *International Journal of Modern Education and Computer Science*, 10(12), 44–53. doi:10.5815/ijmecs.2018.12.06
- Gupta, V. (2013). A Survey of Text Summarizers for Indian Languages and Comparison of their Performance. *Journal of Emerging Technologies in Web Intelligence*, 5(4), 361–366. doi:10.4304/jetwi.5.4.361-366
- Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), 258–268. doi:10.4304/jetwi.2.3.258-268
- Haque, M. M., Pervin, S., & Begum, Z. (2013a). Literature Review of Automatic Multiple Documents Text Summarization. *International Journal of Innovation and Applied Studies*, 3(1), 121–129.
- Haque, M. M., Pervin, S., & Begum, Z. (2013b). Literature Review of Automatic Single Document Text Summarization Using NLP. *International Journal of Innovation and Applied Studies*, 3(3), 857–865.
- Haque, M. M., Pervin, S., & Begum, Z. (2015). Automatic Bengali news documents summarization by introducing sentence frequency and clustering. In *18th International Conference on Computer and Information Technology (ICCIIT)*, (pp. 156 – 160). doi:10.1109/ICCIITech.2015.7488060
- Haque, M. M., Pervin, S., & Begum, Z. (2016). Enhancement of Keyphrase-Based Approach of Automatic Bangla Text Summarization. *Tencon Conference*. doi:10.1109/TENCON.2016.7847955
- Haque, M. M., Pervin, S., & Begum, Z. (2017a). An Innovative Approach of Bangla Text Summarization by Introducing Pronoun Replacement and Improved Sentence Ranking. *Journal of Information Processing Systems*, 13(4), 752–777. doi:10.3745/JIPS.04.0038
- Haque, M. M., Pervin, S., & Begum, Z. (2017b). Rule Based Replacement of Pronoun by Corresponding Noun for Bangla News Documents. *International Journal of Technology Diffusion*, 8(2), 26–42. doi:10.4018/IJTD.2017040102

- Hovy, E. (2005). Automated Text Summarization. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 583–598). Oxford University Press.
- Islam, M. T., & Masum, S. M. A. (2004). Bhasa: A Corpus-Based Information Retrieval and Summariser for Bengali Text. *Proceedings of the 7th International Conference on Computer and Information Technology*.
- Kallimani, J. S., Srinivasa, K. G., & Reddy, E. B. (2014). A Comprehensive Analysis of Guided Abstractive Text Summarization. *International Journal of Computer Science Issues*, 11(6).
- Karim, M. A., Kaykobad, M., & Murshed, M. (2013). *Technical Challenges and Design Issues in Bangla Language Processing*. IGI Global. doi:10.4018/978-1-4666-3970-6
- Kumar, Y. J., & Salim, N. (2012). Automatic Multi Document Summarization Approaches. *Journal of Computational Science*, 8(1), 133–140. doi:10.3844/jcssp.2012.133.140
- Lin, C., & Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. *Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003)*. doi:10.3115/1073445.1073465
- Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2), 159–165. doi:10.1147/rd.22.0159
- Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., & Sundheim, B. (2002). SUMMAC: A text summarization evaluation. *Natural Language Engineering*, 8(1), 43–68. doi:10.1017/S1351324901002741
- Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39–41. doi:10.1145/219717.219748
- Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data* (pp. 43–76). Springer. doi:10.1007/978-1-4614-3223-4_3
- ROUGE 2.0. (n.d.). *Java Package for Evaluation of Summarization Tasks with Updated ROUGE Measures*. Retrieved from <http://kavita-ganesan.com/content/rouge-2.0>
- Saggion, H., & Poibeau, T. (2013). *Automatic Text Summarization: Past, Present and Future*. In *Multi-source, Multilingual Information Extraction and Summarization* (pp. 3–21). Berlin: Springer-Verlag. doi:10.1007/978-3-642-28569-1_1
- Sarkar, K. (2012a). Bengali text summarization by sentence extraction. In *Proceedings of International Conference on Business and Information Management* (pp. 233–245). NIT Durgapur.
- Sarkar, K. (2012b). An approach to summarizing Bengali news documents. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, 857–862. doi:10.1145/2345396.2345535
- Sarkar, K. (2013). Automatic Single Document Text Summarization Using Key Concepts in Documents. *Journal of Information Processing Systems*, 9(4), 602–620. doi:10.3745/JIPS.2013.9.4.602
- Sarkar, K. (2014). A Keyphrase-Based Approach to Text Summarization for English and Bengali Documents. *International Journal of Technology Diffusion*, 5(2), 28–38. doi:10.4018/ijtd.2014040103
- Second most spoken languages around the world. (n.d.). Retrieved August 20, 2015, from <http://graduate.olivet.edu/news-events/news/second-most-spoken-languages-around-world>
- Sikder, R., Hossain, M. M., & Robi, R. H. (2019). Automatic Text Summarization For Bengali Language Including Grammatical Analysis. *International Journal of Scientific & Technology Research*, 8(6), 288–292.
- Uddin, M. A., Sultana, K. Z., & Alam, M. A. (2014). A Multi-Document Text Summarization for Bengali Language. In *The 9th International Forum on Strategic Technology (IFOST)*. Chittagong University of Engineering & Technology (CUET).
- Uddin, M. N., & Khan, S. A. (2007). A Study on Text Summarization Techniques and Implement Few of Them for Bangla Language. *10th International conference on Computer and Information technology*, 1–4. doi:10.1109/ICCITECHN.2007.4579374
- Vector space retrieval model. (n.d.). Retrieved May 10, 2016, from <http://www.ccs.neu.edu/home/jaa/CSG339.06F/Lectures/vector.pdf>

- Words S. Indian Statistical Institute. (2016). *List of stop words for Bengali language*. Retrieved from <http://www.isical.ac.in>
- Yang, L., Cai, X., Zhang, Y., & Shi, P. (2014). Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization. *Information Sciences*, 37-50.
- Ye, S., Chua, T., Kan, M., & Qiu, L. (2007). Document concept lattice for text understanding and summarization. *Journal of Information Process Management, Elsevier*, 43(6), 1643–1662. doi:10.1016/j.ipm.2007.03.010
- Zaman, N. U. (2008). Big Picture Seminar Series. *University of Rochester*. Retrieved December 29, 2015, from <https://www.cs.rochester.edu/u/naushad/survey/BigPicture-URCS-NZ-Bangla.pdf>

Md. Majharul Haque has completed their BSc. Hons in CSE, MS in IT and achieved Doctorate Degree from the CSE Dept of the University of Dhaka. Now, continuing job as Assistant Systems Analyst (Deputy Director) in the Central Bank of Bangladesh.

Anowar Hossain is a passionate software professional with a spur to accept new challenges. Always eager to learn new mechanisms to make my contribution more meaningful.