# An Automated Bengali Text Summarization Technique Using Lexicon-Based Approach

**Busrat Jahan, Sheikh Shahparan Mahtab, Md. Faizul Huq Arif, Ismail Siddiqi Emon, Sharmin Akter Milu, and Md. Julfiker Raju**

**Abstract** There is enough resources for English to process and obtain summarize documents. But this thing is not directly applicable for Bengali language as there is lots of complexity in Bengali, which is not same to English in the context of grammar and sentence structure. Again, doing this for Bengali is harder as there is no established tool to facilitate research work. But this necessary as 26 crore people use this language. So, we have gone for a new approach Bengali document summarization. Here, the system design has been completed by preprocessing the i/p (input) doc, tagging the word, replacing pronoun, sentence ranking, respectively. Pronoun replacement has been added here to minimize the rate of swinging pronoun in the output summary. As the pronoun replacement, we have gone ranking sentences according to sentence frequency, numerical figures (both in digit and word version) and document title. Here, if the sentence has any word that exists in title also taken into our account. The similarity between two sentences has been checked to deduct one as that causes less redundancy. The numerical figure also makes an impact, so they were also identified. We have taken over 3000 newspaper and books documents words has been trained according to grammar. And two

B. Jahan · I. S. Emon · Md. Julfiker Raju
Department of CSE, Feni University, Feni, Bangladesh
e-mail: hossenbipasa980@gmail.com

I. S. Emon
e-mail: emonsahriar0@gmail.com

Md. Julfiker Raju
e-mail: julfikerar@gmail.com

S. S. Mahtab (✉)
Department of EEE, Feni University, Feni, Chittagong Division, Bangladesh
e-mail: mahtabshahzad@gmail.com

Md. Faizul Huq Arif
Department of ICT(DoICT), ICT Division, Dhaka, Bangladesh
e-mail: arifict27@gmail.com

S. A. Milu
Department of CSTE, Noakhali Science and Technology University, Noakhali, Bangladesh
e-mail: sharminmilu7@gmail.com

H. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Lecture Notes in Networks and Systems 171,
https://doi.org/10.1007/978-981-33-4543-0_39

documents have been checked by the design system to evaluate the efficiency of designed summarizer. From the evaluation system, it is been found that the recall, precision, F-score are 0.70 as it is 70%, 0.82 as it is 82%, 0.74 as it is 74%, respectively.

**Keywords** Text summarizer · BTS · Bengali · NLP · Python · Machine learning · POS tagging

## 1 Introduction

Text summarization is the process of summarizing a text or document. There are many summarization tools for the English language. There are also some tasks for automated Bengali text or document summarization. From an application standpoint, the tools do not seem to be very suitable. The abstracts are categorized in two ways: the extractive and the abstractive approach. Most of the summarizer methods for Bengali text summarization are extractive [1]. In an automated text summarization process, a text is delivered to the computer, and the computer returns a less-than-redundant extract or abstract of the original text (s). Text abstraction is the process of producing an abstract or a summary of an extract by selecting a significant portion of the information from one or more texts [1–3]. Thus, the overview summarizes the meaning of the extremes, and some time extraction results in data loss. These methods are not also able to create a plain text from related hierarchical texts. Extract summarization is less like complexity when it comes to favorite issues than abstracts for less complexity. We can use the grammatical rules in conjunction along with mathematical rules for making sentences to decrease the unnecessary error. Again, it can be use  for creating new and plain text from multiple texts that enables to reduce the size of the text summary [4]. Rafel et al. told the extractive summarizer states all the basic requirements. This method has three structures: text analysis, ranking/scoring sentence and summarization [5].

## 2 Literature Review

The observation of summarization of Bangla language for only a document is showed in this sector. However, the area of Bangla text summarization was begun several years back as a new research. Previously, most of the work in the text summarization domain was done on the basis of sentence prohibition. The survey of different text summarization techniques is proposed as the article method [1]. They accomplished an analysis of various methods for text and implemented the basis of extraction Bangla text summarizer.

According to the proposed method of Jones [4], which provides a summary of a text without reading full text we have found that. The main steps in his method have (i) preprocessing, (ii) scoring/ranking sentence and (iii) generating summary. It has also term frequency (TF), inverse document frequency (IDF) and positional value (PV).

The presented method of Haque et al. [5], it summarized Bangla document by using an extraction based summarization technique. The four major steps of their method are given here: (i) preprocessing, (ii) scoring/ranking sentence, (ii) sentence clustering, (iv) generating summary.

Efat et al. [6] suggested a summarization method as an extraction based which acts on the Bangla documents. At the same time, it is capable of summarizing a single document. It has two major steps in their proposed method: (i) preprocessing, (ii) scoring/ranking sentence and summarization.

The method of Das and Bandyopadhyay [7] presented the identification of sentiment from the text, combines it and lastly signifies the text summarization. They used a sentiment model to restore and integrate sentiment. The integration is based on the presentation of theme clustering (K-means) and document level theme relational graph algorithms and finally generates summary selected by the standard page rank algorithm for data retrieval.

# 3 Suggested Method

For successfully we have employed two tagging systems. One is general tagging system, and another is special tagging system. The special tagging system makes the thing best and updated.

## *3.1 General Tagging*

Every word is made to tag (like noun, pronoun, adjective, verb, preposition, etc.). By using a lexicon database [2] and SentiWordNet [3]. The lexicon database and SentiWordNet have limited number of predefined words. Using lexicon database, the words can be tagged as “JJ” (adjective), “NP” (proper noun), “VM” (verb), “NC” (common noun), “PPR” (pronoun), etc. On the other hand, SentiWordNet has list of words with tag as “a” (adjective), “n” (noun), “r” (adverb), “v” (verb), “u” (unknown). Based on these predefined lists of words, we have experimented on 200 Bangla news documents and found that 70% words can be tagged. Bangla words (especially verb) are very much interesting [1]. Though we use word stemming to identify the original term of the word, 100% inactive verbs cannot be stemmed. In fact, it is very difficult to identifying verb because there are many suffixes in Bangla. For example, basis on the tense and person, the English words “do” may be “doing”, “did” and “does”, but on the other hand, the word may have different forms in Bangla. To consider the present continuous tense Like, “কর” (kor-do), three main forms of this word can only depend on the first, second and third person.

Also, it can be “করছি” (doing) for first person, “করছ” (doing) for second person and “করছেন” (doing) for third person, respectively. To consider the present continuous tense Like, “কর” (kor-do), three main forms of this word can only depend on the first, second and third person. Also, it can be “করছি” (doing) for first person, “করছ” (doing) for second person and “করছেন” (doing) for third person, respectively. The forms of verbs for all these meanings of “you” in Bangla are also different. For instance, all these meanings for the forms of verbs of “you” are also different in Bangla. As, “আপনি করছেন” (you are doing), “তুমি করছ” (you are doing), “তুই করছিস” (you are doing) where those terms are specified in present continuous tense and also with second person. Thus, the word “কর” (do) may have the given forms: “করে” (do), “করেন” (do), “করিস” (do), “করি” (do), “করছে” (doing), “করছেন” (doing), “করছ” (doing), “করছিস” (doing), “করছি” (doing), “করেছে” (did), “করেছেন” (did), “করেছ” (did), “করেছিস” (did), “করেছি” (did), “করুক” (do), “করুন” (do), “করল” (did), “করলেন” (did), “করলে” (did), “করলি” (did), “করলাম” (did), “করত” (do), “করতেন” (did), “করতে” (did), “করতিস” (did), “করতাম” (did), “করতেছি” (doing), “করতেছ” (doing), “করতেছেন” (doing), “করছিলি” (doing), “করছিলেন” (doing), “করছিলে” (doing), “করছিলি” (doing), “করছিলাম” (doing), “করেছিলি” (doing), “করেছিলেন” (doing), “করেছিলে” (doing), “করেছিলি” (doing), “করেছিলাম” (doing), “করবে” (do), “করবেন” (do), “করবি” (do), “করব” (do), “করো” (do). Thus, there is no any comparison between the complexity of verb in Bangla and English. However, verb identification is very important for language processing because the verb is the main word of a sentence. So, the complexity of verb in Bangla cannot be compared with English. A list of suffixes are considered as for the final checking in following: “ইতেছিস” (itechhis), “তেছিস” (techhis), “ইতিস” (itis), “ইলে” (ile), “ইবি” (ibi), etc. Now, if the word has suffix, it is tagged as a verb. The result of word tagging has been improved from 68.12% (before using the list of suffixes [4]) to 70% (after using the list of suffixes). We get some preliminary tagging in this step, and later, it may be updated in the next steps and also along with certain words will be specifically tagged as acronym, named entity, occupation, etc., in the next step [8–11].

## 3.2 *Special Tagging*

After general tagging, special tagging was introduced to identify the words as acronym, elementary form, numerical figure, repetitive words, name of occupation, organization and places.

1. Examining for English acronym: When the words are formed by the initials of the other words, then it is called acronym such as “ইউএনও” (UNO), “ওআইসি” (OIC), ““ইউএসএ” (USA). For examining these kinds of words, when we can separate these words that like “ইউএনও” (UNO) to match with “ইউ” (U), “এন”, “ও” (O), those are matched every letter of the words. Actually, we can write all English letters in Bangla like: A for (“এ”), B for (“বি”), C for (“সি”), D for

(“ডি”), … W for (“ডাব্লিউ”), X for (“এক্স”), Y for (“ওয়াই”), Z for (“জেড”) and if we can sort them by descending order depend on their string lengths where W (“ডাব্লিউ”) will be in the first place and A (“এ”) will be in the last place, then match every letter of the words. It is important in descending order that is always used to ensure the longest match. Such as, “এম” (M) does not match with “এ” (A), but it will match with “এম” (M). This experiment shows that 98% success rate for this case.

2. Studying for Bangla elementary tag: Bangla letters with spaces, like: “আ ক ম” (A K M), “এ বি ম” (A B M), etc. These letters will be tagged as Bangla primary tag. We have gotten based on research; the accuracy of the elementary result is 100%.
3. Studying for recurrent words: Recurrent words are special form of word combination where same word can be placed for two times consecutively. For example, “ঠান্ডাঠান্ডা” (thandathanda—cold cold), “বড়বড়” (boroboro—big big), “ছোটছোট” (chotochoto—small small), etc. There are some words, and they are partially repeated such as “খাওয়াদাওয়া” (khawadawa—eat). We have found 100% accuracy on identifying recurrent/repetitive words.
4. Studying for numerical digit: There are three conditions for recognizing the numerical representation in words and digits, are examined as follows:

   (a) It is formed by following the first part of the word, like as, 0 for (০), 1 for (১), 2 for (২),…, 9 for (৯) or “এক” (one), “দুই” (two), “তিন” (three), “চার” (four) to “নিরানব্বই” (ninety nine). The decimal point (.) is also considered when examining the numerical form from digits.
   (b) The next part (if any) is followed by: “শত” (hundred), “হাজার” (thousand), etc.
   (c) Finally, it can have suffixes such as, “টি” (this), “টা” (this), “এন” (en), etc. After the experiment on our sample test documents, 100% numerical form can be found from both numerical values and text documents.

5. Studying for name of occupation: Occupation has a significant word, and for the human named entity identification, occupation is very much helpful by which named entity can be recognized. If we get any word as occupation, we may consider the immediate next some words to find out named entity. We have retrieved some entries for the occupation of Bangladesh from a table such as “শিক্ষক” (shikkhok-master), “সাংবাদিক” (sangbadik-journalist). Every word has matched with these words (that we collected from different online source) and if any matches are found then tagged as occupation. Here, “শিক্ষক” (shikkhok-master) will turn into “প্রধান শিক্ষক” (prodhanshikkhok-Head master) and so on. From this study, it may identify 96% for occupation.
6. Studying for the name of organization: Name of organization is an important factor where any type of word may be the element of organizational name. From our analysis, it has been mentioned as follows:

   (a) The following complete name of the organization, which is depended on the acronym of the name that is together with this parenthesis. For example,

“দূর্নীতদিমনকমশিন(দূদক)” “Durniti Domon Commission (DUDOK) Anti Corruption Commission (ACC)”.

(b) The organization name with last part may contain certain words. Such as, “লমিটিডে” (limited-limited), “বদি্যালয়” (biddaloyschool), “মন্ত্রণালয়” (montronaloy-ministry), etc. Along with the above point, if any such of words are presented in the text according to the point (b), then immediately check the three words of the particular word. Uncertainty when the words are found as noun, name entity or any blocked word, then call them the organizations name. It is found that the organizations name may be accepted the basis of point (b) 85% times.[13–20]

7. Studying for name of place: There is a table the name of places of Bangladesh, it is made with 800 names for the list of division, district, upazila and municipality. Here, the top level is division, second level is district, and third level is upazila or municipality in area-based separation. In addition, we have analyzed 230 countries names and their capitals. In this way, about 91% of the place names can be identified in our experiment.

## 4 Experimental Results

**Sample input**

Title: দুই ভাই-বোনরে ময়না তদন্ত হয়ছে,ে মামলা হয়নি

Text: রাজধানীরবনশ্রীতদেুইভাইবোনরেরহস্যজনকমৃত্যুরঘটনায়এখনোমামলাহয়নি।শিশুদরেবাবামামলাকরবনেবলজোনয়িছেপেরবিার।দুইশিশুরলাশরেময়নাতদন্তহয়ছে।েতাঁদরেগ্রামরেবাড়জিামালপুরলোশদাফনকরাহব।েখাবাররেনমুনাপরীক্ষারফলাফলএখনোপাওয়াযায়নি।শিশুদরে বাবা আমানউল্লাহর বন্ধু জাহদিুল ইসলাম আজ মঙ্গলবার বলো সোয়া ১১ টার দকিে প্রথম আলোকে এসব কথা জানয়িছেনে।রামপুরাথানারভারপ্রাপ্তকর্মকর্তা (ওসি) রফকিুল ইসলাম বলনে, এখনো মামলা হয়নি।পরবিাররে পক্ষ থকেে আজ মামলা হতপোর।েজজি্ঞাসা বাদরে জন্য চায়নজি রস্তেোরাঁর ব্যবস্থাপক, কর্মচারী, পাচককে থানায় নওেয়া হয়ছে।েচায়নজি রস্তেোরাঁ থকেে আগরে দনি আনা খাবার গতকাল সোমবার দুপুরে গরম করে খয়েে ঘুময়িে পড়ে নুসরাত আমান (১২) ও আলভী আমান (৬)। এরপর তারা আর জগেে ওঠনে।ি অচতেন অবস্থায় হাসপাতালে নওেয়া হলে চকিৎিসকরো তাদরে মৃত ঘোষণা করনে।পরবিাররে অভযিোগরে ভত্তিতিে পুলশি জজি্ঞাসাবাদরে জন্য ওই রস্তেোরাঁর মালকিকে থানায় নয়িে গছে।ে নুসরাত ভকিারুননসিা নূন স্কুল অ্যান্ড কলজেরে পঞ্চম ও আলভী হলক্রিসন্টে স্কুলে নার্সারি শ্রণেরি শক্ষির্থী। তাদরে বাবা মো. আমান উল্লাহ ব্যবসায়ী ও মা জসেমনি আক্তার গৃহণিী। এই দম্পতরি এই দুটি সন্তানই ছলি। চায়নজি রস্তেোরাঁ থকেে আগরে দনি আনা খাবার গতকাল সোমবার দুপুরে গরম করে খয়েে ঘুময়িে পড়ে নুসরাত আমান(১২) ও আলভী আমান(৬)। এরপর তারা আর জগেে ওঠনে।ি অচতেন অবস্থায় হাসপাতালে নওেয়া হলে চকিৎিসকরো তাদরে মৃত ঘোষণা করনে। পরবিাররে অভযিোগরে ভত্তিতিে পুলশি জজি্ঞাসাবাদরে জন্য ওই রস্তেোরাঁর মালকিকে ওই দনিই থানায় নয়িে গছে।ে

**Getting Summary of Sample**

Title: দুইশিশুরলাশরেময়নাতদন্তহয়ছে।

Text:রামপুরাথানারভারপ্রাপ্তকর্মকর্তা (ওসি) রফকিুলইসলামবলনে, এখনোমামলাহয়নি।

দুইভাইবোনরেময়নাতদন্তহয়ছে,মামলাহয়নরািজধানীরবনশ্রীতদেুইভাইবোনরেরহস্যজনকমৃত্যুরঘটনায়এখনোমামলাহয়নি।শিশুদরেবাবামামলাকরবনেবলজোনয়িছেপেরবিার। পরবিাররেপক্ষথকেআেজমামলাহতপোর।ে শিশুদরে বাবা আমানউল্লাহর বন্ধু জাহদিুল ইসলাম আজ মঙ্গলবার বলো সোয়া ১১টার দকিে প্রথম আলোকে এসব কথা জানায় ।

See Figs. 1 and 2.

## 4.1 Co-selection Measures

Co-selection measures: In co-selection measures, the principal evaluation metrics are [12]:

(i) **Precision (P)**:

It is the number of sentences occurring in both system generated summary and ideal summary divided by the number of sentences in the system generated summary.

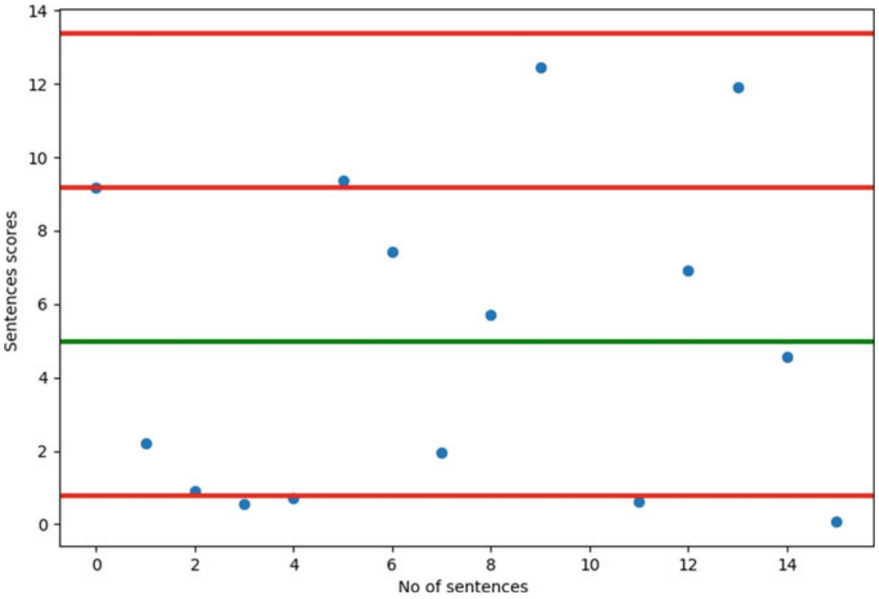

**Fig. 1** Sentence scoring of sample document

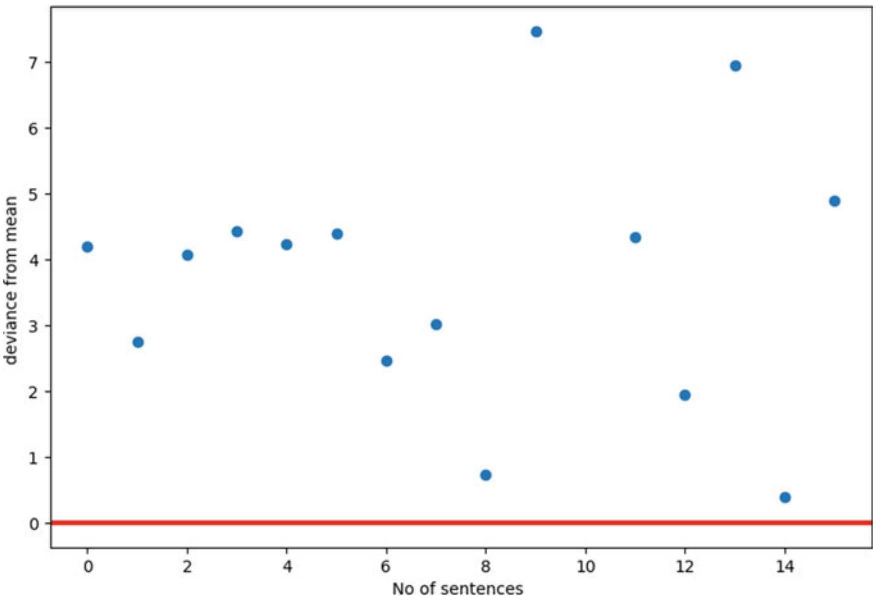

**Fig. 2** Mean deviance of sample document

$$\text{Precision}(P) = (A \cap B)/A$$

where “A” denotes that the number of sentences obtained by the summarizer and also “B” denotes the number of relevant sentences compared to target sets.

(ii) **Recall (R)**:

It is the number of sentences occurring in both systems generated summary and ideal summary divided by the number of sentences in the ideal summary.

$$\text{Recall}(R) = (A \cap B)/B$$

where “A” denotes that the number of sentences obtained by the summarizer and also “B” denotes the number of relevant sentences compared to target sets.

(iii) **F-measure**:

The integrated measure that incorporated both precision and recall is F-measure.

$$\text{F}-\text{Score} = (2 \times P \times R)/(P + R)$$

where “A” denotes that the number of sentences obtained by the summarizer and also “B” denotes the number of relevant sentences compared to target sets.

The evaluation result of first ten document has given in Table 1.

**Table 1** Result of precision, recall and F-score

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

## 5 Conclusion

We have gone for an automatic Bengali document summarizer using Python as a programming platform. There is enough resources for English to process and obtain summarize documents. But this thing is not directly applicable for Bengali language as there is lots of complexity in Bengali, which is not same to English in the context of grammar and sentence structure. Again, doing this for Bengali is harder as there is no established tool to facilitate research work. But this necessary as 26 crore people use this language. So, we have gone for a new approach Bengali document summarization. Here, the system design has been completed by preprocessing the i/p (input) doc, tagging the word, replacing pronoun, sentence ranking, respectively. Pronoun replacement has been added here to minimize the rate of swinging pronoun in the output summary. As the pronoun replacement, we have gone ranking sentences according to sentence frequency, numerical figures (both in digit and word version) and document title. Here, if the sentence has any word that exists in title also taken into our account. The similarity between two sentences has been checked to deduct one as that causes less redundancy. The numerical figure also makes an impact, so they were also identified. We have taken over 3000 newspaper and books documents words which has been trained according to grammar. And two documents have been checked by the design system to evaluate the efficiency of designed summarizer. From the evaluation system, it is been found that the recall, precision, F-score are 0.70 as it is 70%, 0.82 as it is 82%, 0.74 as it is 74%, respectively.

## References

1. Radev, D.R., Hovy, E., McKeown, K.: Introduction to the special issue on summarization. J. Comput. Linguist. **28**(4), 399–408 (2002)
2. Hamou-Lhadj, A., Lethbridge, T.: Summarizing the content of large traces to facilitate the understanding of the behaviour of a software system. In: Proceedings of the 14th IEEE International Conference on Program Comprehension (ICPC), pp. 181–190. IEEE, (2006)
3. Hovy, E.: Automated text summarization. In: Mitkov, R. (ed.) The Oxford Handbook of Computational Linguistics, pp. 583–598. Oxford University Press (2005)
4. Jones, K.S.: Automatic summarizing: factors and directions. In: Advances in Automatic Text Summarization, pp. 1–12 (1999)
5. https://blog.frase.io/
6. Dongmei, A., Yuchao, Z., Dezheng, Z.: Automatic text summarization based on latent semantic indexing. J. Artif. Life Robot. **15**(1), 25–29 (2010)
7. Kunder, M.D.: The size of the world wide web. Online. Available. http://www.worldwidewebsize.com. Accessed 15 Feb 2015
8. Chakma, R., et al.: Navigation and tracking of AGV in ware house via wireless sensor network. In: 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC), Beijing, China, pp. 1686–1690 (2019). https://doi.org/10.1109/cieec47146.2019.cieec-2019589
9. Emon, I.S., Ahmed, S.S., Milu, S.A., Mahtab, S.S.: Sentiment analysis of bengali online reviews written with english letter using machine learning approaches. In: Proceedings of the 6th International Conference on Networking, Systems and Security (NSysS '19). Association for Computing Machinery, New York, pp. 109–115 (2019). doi: https://doi.org/10.1145/3362966.3362977
10. Ahmed, S.S., et al.: Opinion mining of Bengali review written with English character using machine learning approaches. In: Bindhu V., Chen J., Tavares J. (eds.) International Conference on Communication, Computing and Electronics Systems. Lecture Notes in Electrical Engineering, vol. 637. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2612-1_5
11. Milu, S.A., et al.: Sentiment Analysis of Bengali reviews for data and knowledge engineering: a Bengali language processing approach. In: Bindhu V., Chen J., Tavares J. (eds.) International Conference on Communication, Computing and Electronics Systems. Lecture Notes in Electrical Engineering, vol. 637. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2612-1_8
12. Munir, C., Ibrahim, K., Mofazzal, H.C.: Bangla VasarByakaran. Ideal publication, Dhaka (2000)
13. Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R.D., de Frana Silva, G., Simske, S.J., Favaro, L.: A multi-document summarization system based on statistics and linguistic treatment. Expert Syst. Appl. **41**(13), 5780–5787 (2014)
14. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
15. Foong, O.M., Oxley, A., Sulaiman, S.: Challenges and trends of automatic text summarization. Int. J. Inf. Telecommun. Technol. **1**(1), 34–39 (2010)
16. Azmi, A.M., Al-Thanyyan, S.: A text summarizer for arabic. J. Comput. Speech Lang. **26**(4), 260–273 (2012)
17. Karim, M.A., Kaykobad, M., Murshed, M.: Technical challenges and design issues in bangla language processing. Published in the United States of America by Information Science Reference (an imprint of IGI Global) (2013)
18. Islam, M.T., Masum, S.: Bhasa: a corpus based information retrieval and summarizer for bengali text. In: Proceedings of the 7th International Conference on Computer and Information Technology (2004)

19. Uddin, M.N., Khan, S.A.: A study on text summarization techniques and implement few of them for bangla language. In: Proceedings of the 10th International Conference on Computer and Information Technology (ICCIT-2012), pp. 1–4. IEEE (2007)
20. Sarkar, K.: Bengali text summarization by sentence extraction. In: Proceedings of International Conference on Business and Information Management (ICBIM-2012), pp. 233–245. NIT Durgapur (2012)