An Automated Bengali Text Summarization Technique Using Lexicon Based Approach

Chapter · September 2020

7 authors, including:

Busrat Jahan
Feni University
7 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Sharmin Akter Milu
Noakhali Science & Technology University
7 PUBLICATIONS **4** CITATIONS

SEE PROFILE

s.s Mahtab
Dhaka University of Engineering & Technology
14 PUBLICATIONS **4** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Association Rule View project

Project   Embedded Lab SCUT View project

An Automated Bengali Text Summarization Technique Using Lexicon Based Approach

Busrat Jahan[1], Sheikh Shahparan Mahtab[2], Md. Faizul Huq Arif[3], Ismail Siddiqi Emon[4], Sharmin Akter Milu[5], Md. Julfiker Raju[6]

[1]Department of CSE, Feni University, Feni, Bangladesh
Email: hossenbipasa980@gmail.com
[2]Department of EESE,Universiti Kebangsaan Malaysia , Bangi, Selangor,Malaysia
Email: mahtabshahzad@gmail.com
[3]Department of ICT(DoICT),ICT Division, Bangladesh
Email: arifict27@gmail.com
[4]Department of CSE, Feni University, Feni, Bangladesh
Email: emonsahriar0@gmail.com
[5]Department of CSTE, Noakhali Science and Technology University, Noakhali, Bangladesh
Email: sharminmilu7@gmail.com
[6]Department of CSE, Feni University, Feni, Bangladesh
Email: julfikerar@gmail.com

{Corresponding Author: Sheikh Shahparan Mahtab[2*]
Email: mahtabshahzad@gmail.com}

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), 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 cause's 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 documents has 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 is 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 [3].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. Alternatively, 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 foronly a document is showedin this sector. However the area of Bangla text summarization was begun several yearsback as a new research. Previously, most of the work in the text summarization domain was done on the basisof sentence prohibition. The surveyof different text summarization techniques are proposed as the article method [3]. They accomplished an analysis of variousmethods for text and implemented the basis of extraction Bangla text summarizer.

According to the proposed method of K. Sarkar [4], which providesa summary of a text without reading full text. The main steps in his method has(i) preprocessing, (ii) scoring /ranking sentence and (iii) generatingsummary. It has also term frequency (TF), inverse document frequency (IDF) and Positional Value (PV).

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

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

The method of A. Das and S. Bandyopadhyay [7] presented,the identification of sentiment from the text, combine it and lastly signify the text summarization. They used a sentiment model to restore and integrate sentiment. The integration isbased on the presentation of theme clustering (K-means) and documentlevel 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 system. One is general tagging system and another is Special tagging system. The special tagging system makes the thing best and updated.

3.1General Tagging

Every word is made to tag (like noun, pronoun, adjective, verb, preposition, etc.). by using a lexicon database [1] and SentiWordNet [2]. The lexicon database and SentiWordNet has 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 [3]. Though we use word stemming to identify the originalterm of the word, 100% inactive verbs can't 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 BanglaTo consider the present continuous tense Like, "কর" (kor-do) three main forms of this word can only have 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 have 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, "আপনিকরছেন" (you are doing), "তুমিকরছ" (you are doing), "তুইকরছিস" (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),"করছিল" (doing), "করছিলেন" (doing), "করছিলে"

(doing), "করছিলি" (doing), "করছিলাম" (doing),"করেছিল" (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 can't 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 taggingin this step and later it may be updated in the next steps and also along with certain words will be specifically tagged asacronym, named entity, occupation, etc. in the next step[17-20].

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), etc. 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: " আকম " (A K M), " এবিম " (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, ''ঠাণ্ডাঠাণ্ডা'' (thandathanda - cold cold), "বড়বড়" (boroboro - big big), "ছোটছোট"(chotochoto - small small), etc. There are some words, 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), etc. 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 "প্রধানশিক্ষক"   (prodhanshikkhok-Head master), and so on. From this study, it may identify 96% for occupation.

6.Studying for the name of organizationName 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, ''দুর্নীতিদমনকমিশন(দুদক)''"DurnitiDomonCommission(DUDOK)AntiCorruptionCommission(ACC)''.

(b) The organization name with last part may contain certain words. Such as," লিমিটেড''(limited-limited),"বিদ্যালয়''(biddaloyschool),"মন্ত্রণালয়"(montronaloy-ministry) etc. [93]. 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.

7.Studying for name of place: There is a table thename 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: দুইভাই-বোনেরময়নাতদন্তহয়েছে, মামলাহয়নি

Text:
রাজধানীরবনশ্রীতেদুইভাইবোনেররহস্যজনককমৃত্যুররঘটনায়এখনোমামলাহয়নি।শিশুদেরবাবামামলাকরবেনবলেজানিয়েছেপরিবার।দুইশিশুরলাশেরময়নাতদন্তহয়েছে।তাঁদেরগ্রামেরবাড়িজামালপুরেলাশদাফনকরাহবে।খাবারেরনমুনাপরীক্ষারফলাফলএখনোপাওয়াযায়নি।শিশুদেরবাবাআমানউল্লাহরবস্তুজাহিদুলইসলামআজমঙ্গলবারবেলাসোয়া১১টারদিকেপ্রথমআলোকেএসবকথাজানিয়েছেন।রামপুরাথানারভারপ্রাপ্তকর্মকর্তা   (ওসি)   রফিকুলইসলামবলেন, এখনোমামলাহয়নি।পরিবারেরপক্ষথেকেআজমামলাহতেপারে।জিজ্ঞাসাবাদেরজন্যচায়নিজরেস্তোরাঁরব্যবস্থাপক,   কর্মচারী, পাচককেথানায়নেওয়াহয়েছে।চায়নিজরেস্তোরাঁথেকেআগেরদিনআনাখাবারগতকালসোমবারদুপুরেগরমকরেখেয়েঘুমিয়েপড়েনুসরাতআমান   (১২) ওআলভীআমান(৬)।এরপরতারাআরজেগেওঠেনি।অচেতনঅবস্থায়হাসপাতালেনেওয়াহলেচিকিৎসকেরাতাদেরমৃতঘোষণাকরেন।পরিবারেরঅভিযোগেরভিত্তিতেপুলিশজিজ্ঞাসাবাদেরজন্যওইরেস্তোরাঁরমালিককেথানায়নিয়েগেছে।নুসরাতভিকারুননিসানূনস্কুলঅ্যান্ডকলেজেরপঞ্চমওআলভীহলিক্রিসেন্টস্কুলেনার্সারিশ্রেণিরশিক্ষার্থী।তাদেরবাবামো. আমানউল্লাহব্যবসায়ীওমাজেসমিনআক্তারগৃহিণী।এইদম্পতিরএইদুটিসন্তানইছিল।চায়নিজরেস্তোরাঁথেকেআগেরদিনআনাখাবারগতকালসোমবারদুপুরগরমকরেখেয়েঘুমিয়েপড়েনুসরাতআমান(১২) ওআলভীআমান(৬)।এরপরতারাআরজেগেওঠেনি।অচেতনঅবস্থায়হাসপাতালেনেওয়াহলেচিকিৎসকেরাতাদেরমৃতঘোষণাকরেন।পরিবারেরঅভিযোগেরভিত্তিতেপুলিশজিজ্ঞাসাবাদেরজন্যওইরেস্তোরাঁরমালিককেওইদিনইথানায়নিয়েগেছে।

Getting Summary of Sample

Title: দুইশিশুরলাশেরময়নাতদন্তহয়েছে।

Text:রামপুরাথানারভারপ্রাপ্তকর্মকর্তা (ওসি) রফিকুলইসলামবলেন, এখনোমামলাহয়নি। দুইভাইবোনেরময়নাতদন্তহয়েছে।মামলাহয়নিরাজধানীরবনশ্রীতেদুইভাইবোনেররহস্যজনকমৃ ত্যুরঘটনায়এখনোমামলাহয়নি।শিশুদেরবাবামামলাকরবেনবলেজানিয়েছেপরিবার।পরিবারেরপক্ষ থেকেআজমামলাহতেপারে।শিশুদেরবাবাআমানউল্লাহরবন্ধুজাহিদুলইসলামআজমঙ্গলবারবেলাসো য়া১১টারদিকেপ্রথমআলোকেএসবকথাজানায়।
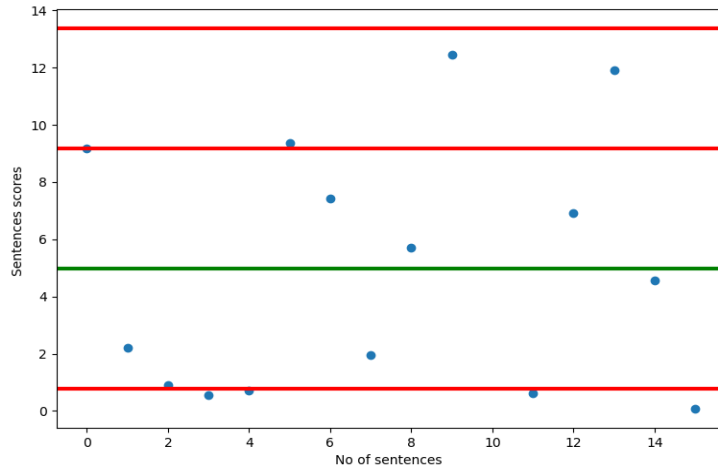

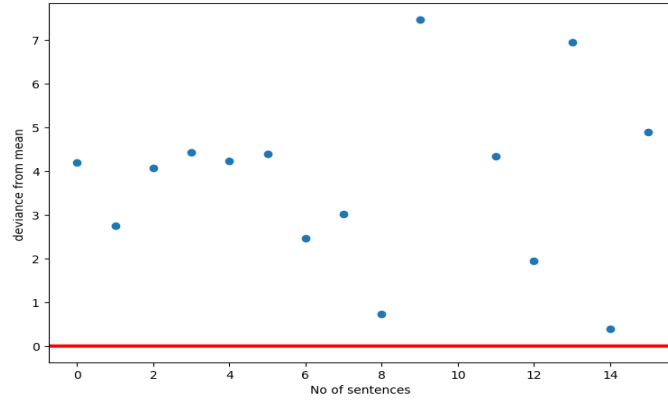**Figure 1.**Sentence Scoring of Sample Document

Figure 2. Mean Deviance of Sample Document

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.
Precision (P) = (A∩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.
Recall(R) = (A∩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.

F-Score = (2×P×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 fist 10 document has given below 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	075
3	0.82	0.69	0.74
4	0.82	0.68	0.74
5	0.79	0.71	074
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	**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), 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 cause's 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 documents has 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 is 0.70 as it is 70%, 0.82 as it is 82%, 0.74 as it is 74% respectively.

References

[1]   A. Hamou-Lhadj and T. Lethbridge, "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). IEEE, 2006, pp. 181–190.

[2]   E. Hovy "Automated Text Summarization". In: Mitkov, R. (Ed.), the Oxford Handbook of Computational Linguistics. Oxford University Press, 2005, chapter 32, pp. 583-598.

[3]   D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," Journal of Computational Linguistics, vol. 28, no. 4, pp. 399–408, December 2002.

[4]   K. S. Jones, "Automatic summarizing: factors and directions," Advances in automatic text summarization, pp. 1–12, 1999.

[5]   [https://blog.frase.io/

[6]   A. Dongmei, Z. Yuchao, and Z. Dezheng, "Automatic text summarization based on latent semantic indexing," Journal of Artificial Life and Robotics. Springer, vol. 15, no. 1, pp. 25–29, August 2010.

[7]   M. d. Kunder, "The size of the world wide web," February 2015, [Online]. Available:http://www.worldwidewebsize.com.[Accessed: 15- February-2015].

[8] R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. de Frana Silva, S. J. Simske, and L. Favaro, "A multi-document summarization system based on statistics and linguistic treatment," Expert Systems with Applications. Elsevier, vol. 41, no. 13, pp. 5780–5787, October 2014.

[9] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research and Development, vol. 2, no. 2, pp. 159–165, April 1958.

[10] [10] O. M. Foong, A. Oxley, and S. Sulaiman, "Challenges and trends of automatic text summarization," International Journal of Information and Telecommunication Technology, vol. 1, no. 1, pp. 34–39, 2010.

[11] A. M. Azmi and S. Al-Thanyyan, "A text summarizer for arabic,"JournalofComputer Speech and Language, vol. 26, no. 4, pp. 260–273, August 2012.

[12] C. Munir, K. Ibrahim, and H. C. Mofazzal, Bangla VasarByakaran. Ideal publication, Dhaka, November 2000.

[13] M. A. Karim, M. Kaykobad, and M. Murshed, "Technical Challenges andDesign Issues in Bangla Language Processing". Published in the United States of America by Information Science Reference (an imprint of IGI Global), June 2013

[14] M. T. Islam and S. Masum, "Bhasa: A corpus based information retrieval and summarizer for bengali text," in Proceedings of the 7th International Conference on Computer and Information Technology, December 2004.

[15] M. N. Uddin and S. A. Khan, "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). IEEE, 2007, pp. 1–4.

[16] K. Sarkar, "Bengali text summarization by sentence extraction," in Proceedings of International Conference on Business and Information Management (ICBIM-2012), NIT Durgapur, 2012, pp. 233–245.

[17] R. Chakma et al., "Navigation and Tracking of AGV in ware house via Wireless Sensor Network," 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC), Beijing, China, 2019, pp. 1686-1690, doi: 10.1109/CIEEC47146.2019.CIEEC-2019589.

[18] Ismail Siddiqi Emon, Sabiha Sunjida Ahmed, Sharmin Akter Milu, and S. S. Mahtab. 2019. 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, NY, USA, 109–115. DOI:https://doi.org/10.1145/3362966.3362977

[19] Ahmed S.S. et al. (2020) 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. https://doi.org/10.1007/978-981-15-2612-1_5

[20] Milu S.A. et al. (2020) 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. https://doi.org/10.1007/978-981-15-2612-1_8