Experimental Evaluation andApproach of Enhancement in Generation of Automatic Unsupervised Extractive Text Summarization of Marathi Text By Using Machine Learning Algorithm

[1]Apurva D. Dhawale, [2]Sonali B. Kulkarni, [3]Vaishali M. Kumbhakarna
[1,2] Dept. of Computer Science & IT,
[1,2,3] Dr. B. A. M. University, Aurangabad, India
[1]addhawale@gmail.com, [2]sonalibkul@gmail.com, [3]vmk_17@yahoo.co.in

ArticleInfo
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202202004
Received 10 December 2021; Revised form 25 December 2021; Accepted 30 December 2021
Available online 05 January 2022.

Abstract - The Text summarization has immense importance in the arena of Natural Language Processing. The summarization can be done in two ways: abstractive and extractive text summarization. Here, the language is a vast query, because we have large number of national and international languages. By studying the literature, the Marathi language is chosen for this work and we are trying to make a framework where all the students giving competent examinations can have summarized Marathi e-news articles by using extractive text summarization. For this objective we have used TextRank algorithm which has proven very effective for different languages. This paper demonstrates the summarization techniques on Marathi e-news articles using Gensim Library of TextRank algorithm and comparative analysis of summaries generated by using both TextRank and ROUGH method.

Keywords - Machine Learning approach, unsupervised method, Gensim, Extractive text summarization, TextRank algorithm, Ratio, Wordcount

I. INTRODUCTION

Text summarization has been performed on various plentiful protuberant languages [1]. Indian languagesdo not have plenty of monolingual corpora which is large and publicly available, but they are broadly spoken bymore than a billion speakers,theyinclude 8out of top 20 utmost spoken languages and ∼30 languageswith more than a million speakers. There isalso anincreasingpopulace of users consuming Indianlanguage content (print, digital, governmentand businesses).[4] The Marathi language is a prominent and predominantly spoken language in Maharashtra, India. But as far as the research is concerned, it has not established the required attention in the domain of Natural Language Processing. This language is morphologically rich.

A lot of study has been done on extractive summarization of text that it works superbly as it doesn't change the construction of a sentence and methods of POS provided by any language. For numerous languages across the world, for python, the TextRank algorithm is used, and for Marathi Text which we are focusing, Gensim library works efficiently. We are creating a framework which summarizes e-news articles in Marathi for the students who are useful for the students preparing for competitive examinations [2].

Our efforts are primarily in this direction to develop acompetent framework which supports students in easiest way for upgrading themselves with recent trends and current awareness by summarized e-news articles. This paper prominences on single document Extractive Marathi text summarizer.

Semantic representation and sentence extraction are very effective in extractive technique compared to abstractive. The monolingual summary here is a competent way of representing the long text in sort form.

II. WORK IN THIS AREA

There are 2 ways to perform text summarization: first one is single document and the other is multidocument, and the summarization further divided into 2 types: 1) extraction method and 2) abstraction method. keywords and paragraphs are inhibited to create summaries. In extraction-based summarization, the assembled innovative text is formed in abstractive summarization [3].

Considering English language, a lot of work has been done and got enhanced results for summarizing it. it can be predicted that there is a need to focus on regional languages in India which are used in numerous fields for numerousdeterminations. A major problem with web contents can be redundancy in text and there is no standard evaluation method for some problems in NLP. Summarization is easier than evaluation considering this problem. [5]

In this study, researchers have projected an enhancedapproachto solve the problem of topic deviation in abstractive summarization method by combining TextRank with BART model. These methods are used to extract and create summaries from news text data. Tentative results show that compared with single BART model, the average recall scores of Rouge-1, Rouge-2 and Rouge-Lare improved by 1.5%, 0.5% and 1.3% respectively. [7]

This paper presents a system for unsupervised extractive text summarization method using neural networks for Marathi e-news articles. Extractive summaries or extracts are formed by detectingsignificant sentences or words which are directly selected from the given input document. They tried to propose the use of atype of neural network i.e., Recurrent Neural Network (RNN) which can perform calculations on sequential data.[8]

This paper focuses on overcoming the scalability problem by extracting tokens in Marathi text [9] diverse datasets are undergone the experiments and results prove thestrength of the proposed method for Marathi Corpus by using the precision and entropy.

This study emphases on lack of resources, either tools or corpora, which is a noticeable problem for both linguistic methods & statistical methods. Machine learning is often used with both statistical and linguistic approaches to train the system using a labelled corpus. since a summary has to be evaluated using many aspects, assessing different methods is challenging.[10]

In this paper, authors proposed a survey on existing text summarization methods and NLP tools for Indian regional languages. The issues associated with the Indian languages are also elaborated which are the bottlenecks for summarization of Indian languages.[11]

III. PROPOSED METHODOLOGY

We have used dataset from GitHub which includes 1135 Marathi e-news articles which are divided into 5 domains as banking, sports, film industry, general knowledge& polity news. The type of the files in dataset is '.txt' and the contents of these files are then forwarded as an input for summarization system.[6]

In this study, we are using extractive based approach using machine learningTextRank algorithm. As shown in above figure, the system reads the document first, then its length is calculated, and it would produce a summary which gives user;significant sentences according to the necessity of the user. The digital textual datais acquiringrank day by day, &the related literature demonstrates that there are many methods & algorithms suitable for Text processing and text summarization. depending on the language elected and the algorithm chosen, the result may differ.

By using sentences as nodes, the TextRank mock-ups any document as a graph for the task of automated text summarization.[14]. The unstructured data is converted in structured form by performing these experiments. The TextRank algorithm trails the mentioned phases to get summary.

The primary foot steps to calculate the total length of the total input Marathi text. The total length is compared with the summarized text length at the end, this is how the user gets the ratio to be tested.

Formerly the word split step is performed byusing mytext.split() function & it is saved in other variable.Each sentence is split into words, the frequency of each word is then counted & stored in an empty array. The frequency count is calculated by using get () function and counter will benefit to get exact count of individual word.

In the next step Key Value pairs are created. Theefficiency of an algorithm is subject to the language used for deriving summary. This system helps to progress theresults by data pre-processing, and improved efficiency can be seen in case of Marathi text.[15]

The meaning of Gensim is to "Generate Similar", it is used for unsupervised topic modelling. Gensim is an open-source library.[13]

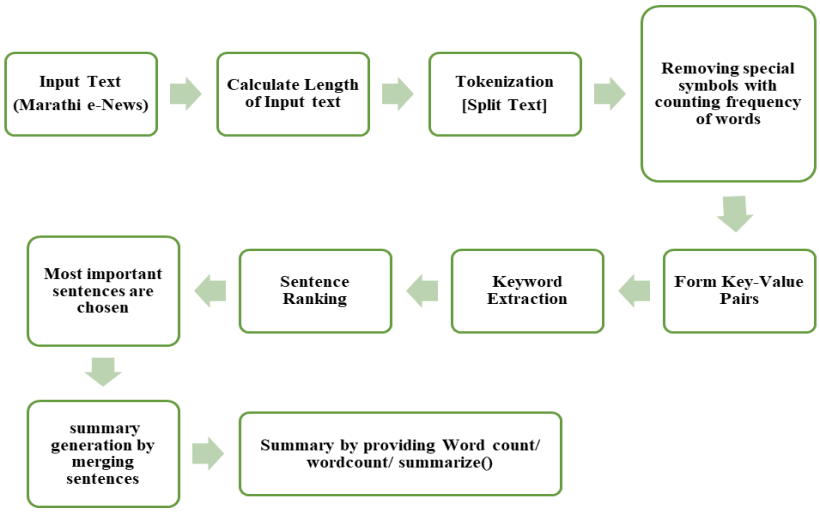

Fig 1. Proposed Methodology

There are 3 methods included in Gensim library of Python, which helps user to generate meaningful summary of input Marathi e-news article. This algorithm provisions unsupervised extractive Marathi text summarization which generates summary by combining top rated sentences from the input data file.

First method is to use Summarize () function from Gensim library using TextRank Algorithm, this method shows the summary of input text by using 0.2% ratio. i.e. 20%. This needs minimum 200 characters in input text.

The second method takes input of wordcount from user which we have considered 150 with which user can have meaningful summary. So, with the 150-character wordcount the summary is generated. This value of count may be increased or decreased as per user requirements.

The Third Method takes ratio value as an input from user which varies between 0 to 1. Here, 0.3 refers to 30% of summary, similarly 0.5 refers to 50% summary which we have considered for our testing news articles. The default value is 0.2.

IV. EXPERIMENTAL ANALYSIS &RESULTS

As explained above the first method shows the 20% summary of input text by using summarize function of Gensim library. The resultant values of lengths of input and output text are arranged in a vector table and shown with the help of graphical representation as follows.

Here in this graph, we have considered 30 samples of banking domain. The values on red bar shows total length and orange bar shows summary length. The first news article b1 is having input text length of 2372 and its length after generation of summary is 460, which is approx. 20% of the original text. It is verified for all other samples too. So, the default value of this function is 0.2.[12]

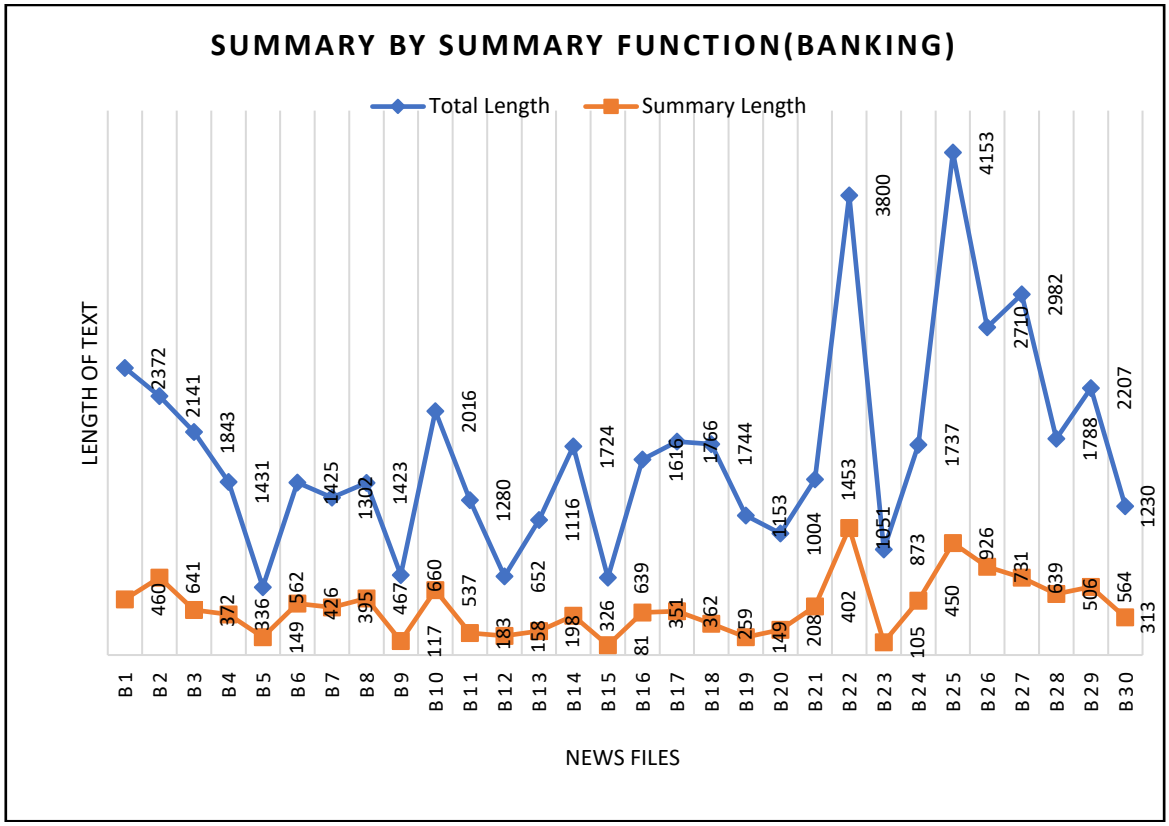

Fig 2. Summary by using summary()

The second method's vector table and graphical representation is as follows:

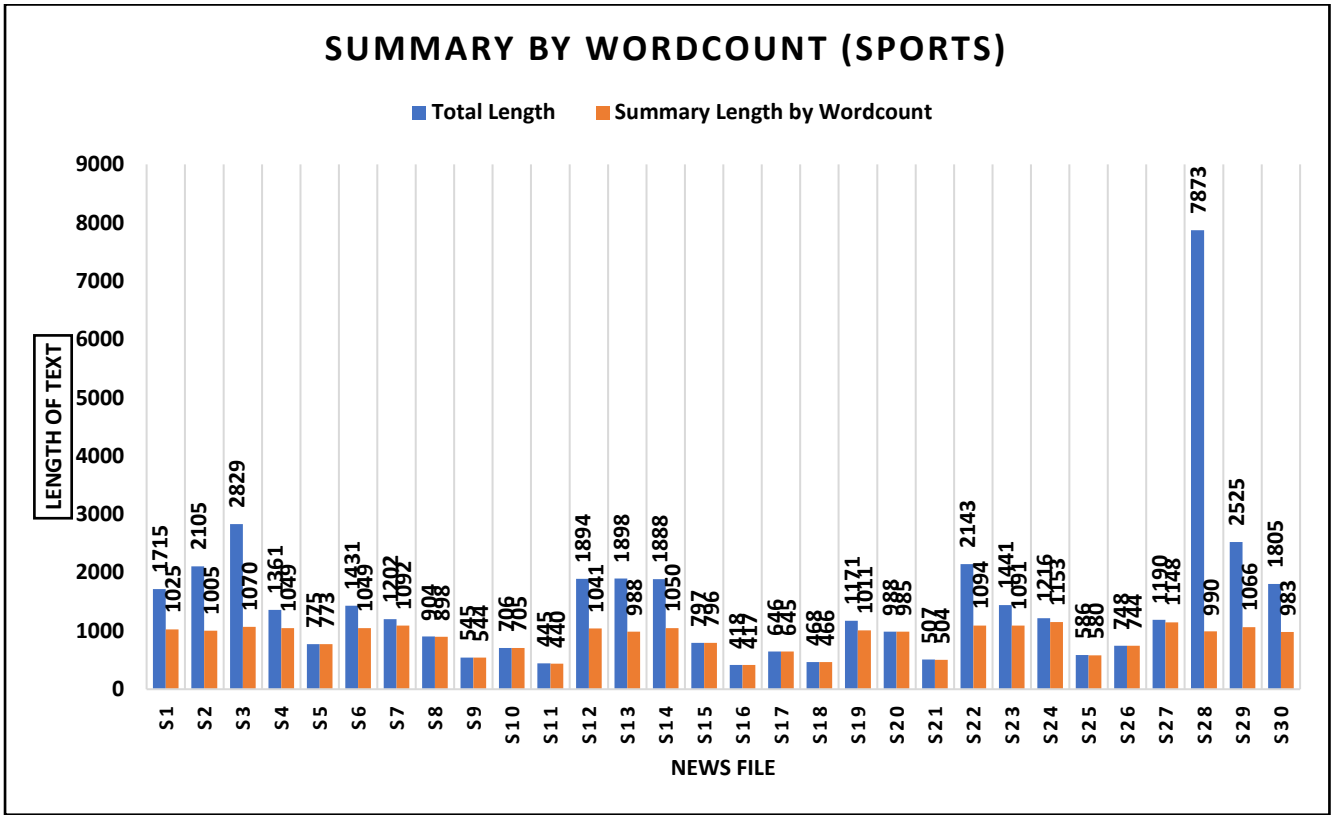

Fig 3. Summary by using Wordcount

The above graph shows the 30 samples chosen from the sports domain where news files are summarized on the basis of wordcount which is considered 150 here for testing purpose. It may vary as per need of the user. The news samples are summarized using Gensim library & the graph shows the accuracy which is further compared with the summarize function. The third method's vector table and graphical representation is as follows:

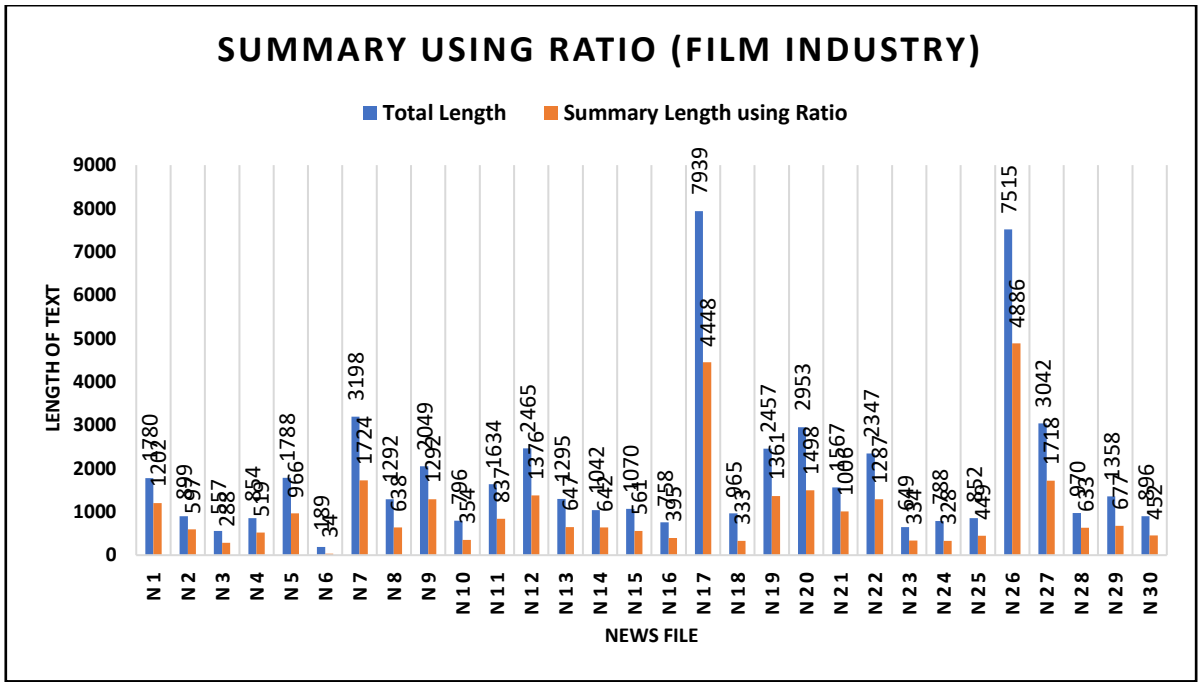

Fig 4. Summary by using Ratio

This graph shows the summarization of 30 samples from film industry domain. Here the summary is generated using ratio. The value of ratio can vary from 0 to 1. We have considered 0.5 for testing purpose. Minimum and default value is 0.2, as it requires meaningful summary. The accuracy of this method is also compared with other 2 methods mentioned above.

Here, we are scheming the summary on the basis of either summary (), wordcount or ratio functions in Gensim Library using TextRank algorithm which works best on the Marathi text. The following graph shows the summary of news in general knowledge domain by using all the 3 methods that are considered at a time, so that the user can go through the comparative summary analysis of same news article by 3 different methods.

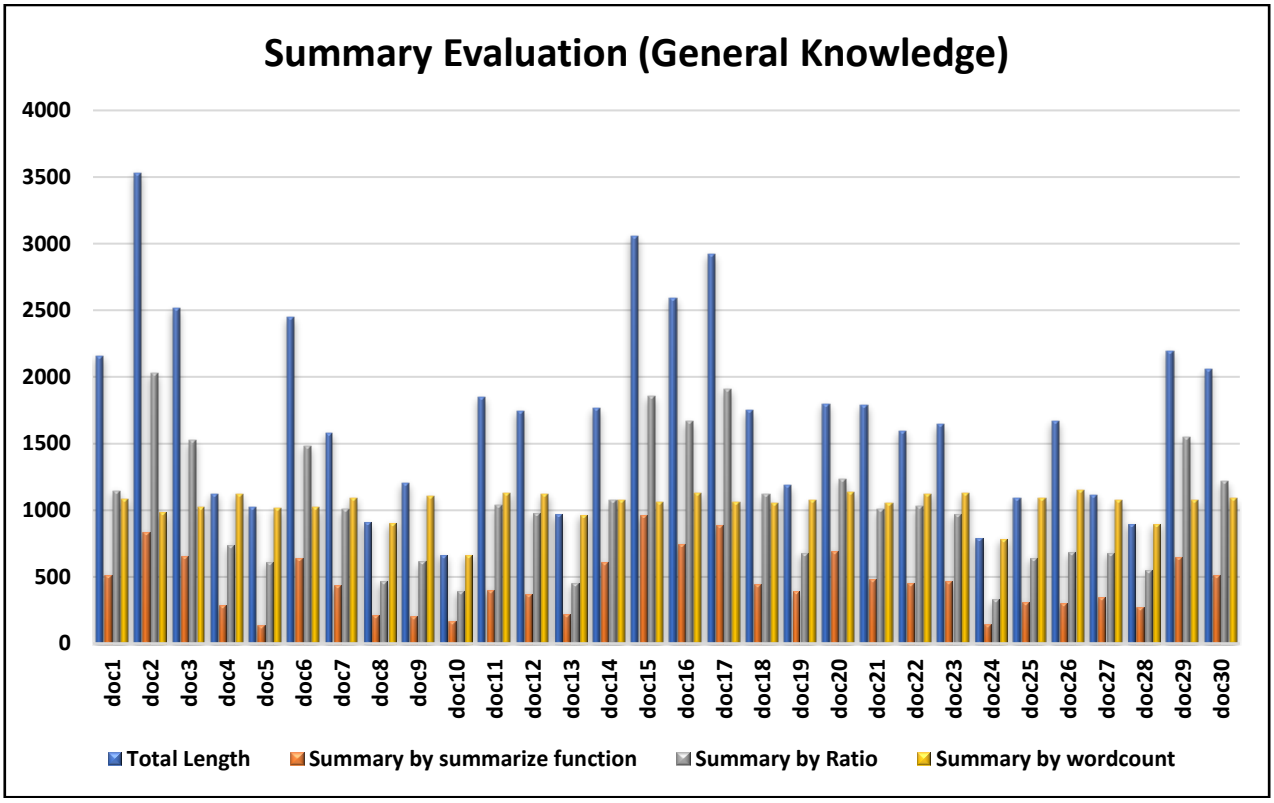

Fig 5. Summary evaluation by all methods(Combined representation)

The successful summary i. e. positives and negatives are also evaluated. The following table shows the count: The percentage of accuracy of individual domain is calculated based on the above table which is shown below:

Table 1. Percentage of Accuracy

Domain	Total	positives	Negatives	Percentage
Banking	30	29	1	96.67
Sports	30	23	7	76.67
Film Industry	30	29	1	96.67
General Knowledge	30	29	1	96.67
Politics	30	24	6	80.00

The following graphs shows the results:

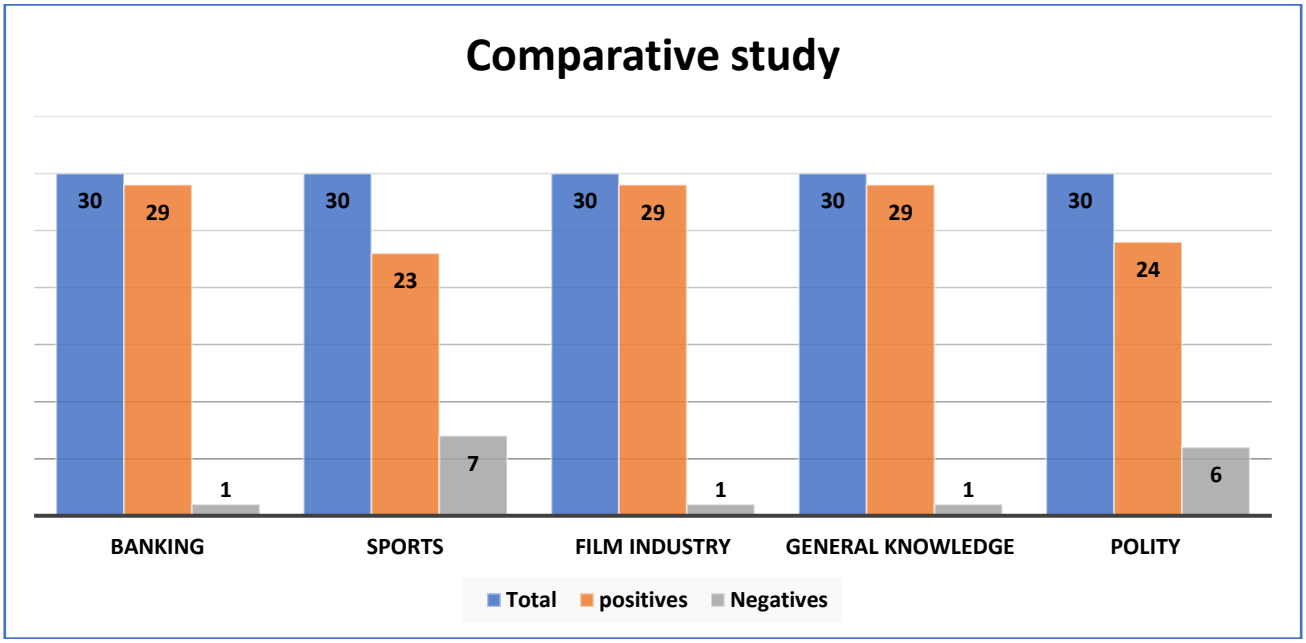

Fig 6. Comparative study

Here in this graph the blue bars show the total samples considered, orange bar shows the true positives i.e. appropriate summary generation, and the Gray bar shows the negatives, i.e. summary is not generated up to the mark. It covers all 5 domains which we have covered.

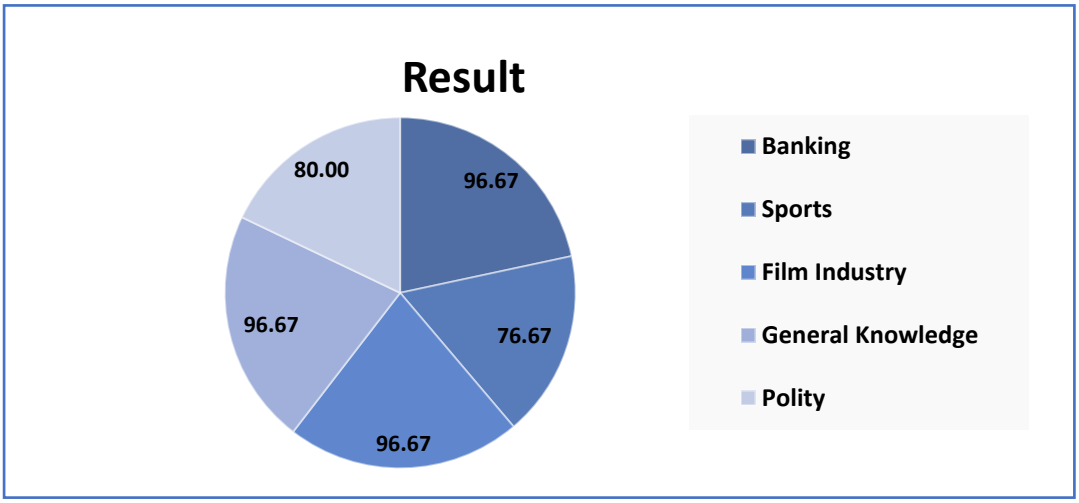

Fig 7. Result Analysis of all 5 domains

This graph shows the results of all domains which are calculated on the basis of the comparative study values of positive and negative summary samples. So the overall average accuracy of all the methods combined is 96.66%. This shows that the system really works well with Marathi text unsupervised extractive summary generation.

V. CONCLUSION

The paper explicates the experiments which are performedon the Marathi e-news of 5 different domains like Banking, Sports, Film Industry, General Knowledge, & Politics. The TextRank algorithm is used for summarizing News articles and the outcomes shows that it works well with Banking, Film Industry, & General Knowledge with 96.67% result for all and sports & politics with 76.67% & 80% respectively. Considering the positives and negatives in the news samples, the results of all three methods are compared and the average accuracy is 96.66%. This summarization technique may prove to be more precise and obliging to the society.

References

[1]. Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna, "Automatic Pre-Processing of Marathi Text for Summarization", International Journal of Engineering and Advanced Technology (IJEAT) Volume-10 Issue-1, October 2020.

[2]. Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna, "Automatic Unsupervised Extractive Summarization of Marathi Text Using Natural Language Processing", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 22, Issue 6, Ser. II, PP 21-25, Nov. – Dec. 2020.

[3]. Dhawale A.D., Kulkarni S.B., Kumbhakarna V.M. (2021)" A Survey of Distinctive Prominence of Automatic Text Summarization Techniques Using Natural Language Processing". In: Raj J.S. (eds) International Conference on Mobile Computing and Sustainable Informatics. ICMCSI 2020.

[4]. Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C.,Avik Bhattacharyya, Mitesh M. Khapra, Pratyush Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages", Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948–4961. Association for Computational Linguistics, November 16 - 20, 2020.

[5]. Dhawale A.D., Kulkarni S.B., Kumbhakarna V.M., "Survey of Progressive Era of Text Summarization for Indian and Foreign Languages Using Natural Language Processing", In: Raj J., Bashar A., Ramson S. (eds) Innovative Data Communication Technologies and Application. ICIDCA 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 46. Springer, Cham.

[6]. https://github.com/chiragsanghvi/TextSummarizer

[7]. Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2021, pp. 2005-2010, doi: 10.1109/IAEAC50856.2021.9390683.

[8]. Anishka Chaudhari, Akash Dole, Deepali Kadam, "Marathi text summarization using neural networks", International Journal of Advance Research and Development (IJARnD), Volume 4, Issue 11,2019.

[9]. Prafulla B. Bafna, Jatinderkumar R. Saini, "Marathi Text Analysis using Unsupervised Learning and Word Cloud", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020.

[10]. Abdelkrime Aries Djamel eddine Zegour Walid Khaled Hidouci, "Automatic text summarization: What has been done and what has to be done", ArXiv Journal, volume abs/1904.00688, 1st April 2019

[11]. Pradeepika Verma, Anshul Verma, "Accountability of NLP Tools in Text Summarization for Indian Languages", Journal of Scientific Research, Institute of Science, Banaras Hindu University, Varanasi, India, Volume 64, Issue 1, 2020.

[12]. https://tedboy.github.io/nlps/generated/generated/gensim.summarization.summarize.html#:~:text=summarize(),-gensim.summarization.&text=Returns%20a%20summarized%20version%20of,be%20given%20as%20a%20string.

[13]. https://www.tutorialspoint.com/gensim/gensim_introduction.htm

[14]. Christopher D. Manning, Prabhakar Raghavan, H.S.: Introduction to Information Retrieval. Cambridge University Press (2008).

[15]. Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna, "A Machine Learning Approach for Automatic Unsupervised Extractive Summarization of Marathi Text", International Journal of Creative Research Thoughts (IJCRT), Volume 8, Issue 11, November 2020.