IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Machine Learning Approach for Automatic Unsupervised Extractive Summarization of Marathi Text

Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna
Dept. of Computer Science & IT, Dr. B. A. M. University, Aurangabad, India.

Abstract: The Data Science is showing us the upward mark in terms of popularity, and efficiency in dealing with e-contents. The Natural Language Processing is playing vital role in e-content data manipulations. The reading of e-contents by the users is mostly preferred in their regional language, where numerous foreign and Indian Regional languages come in lime light. We are proposing a system which will lead students to get up-to-date recent Marathi e-news in summarized way, as it is the regional language of Maharashtra, India. It is observed that very less work has been done in Marathi language text summarization. This paper emphasises on the processing of e-news to get extractive summary. We have followed TextRank algorithm to get summary by providing ratio which will be the value between 0 and 1.

Keywords: TextRank, Automated Text Summarization (ATS), Information Retrieval, Latent Dirichlet Allocation (LDA), LINGO (Label Induction Grouping).

I. INTRODUCTION

The Automated Summarization is a core area of Natural Language Processing system. This aspect is impressive because the e-contents are majorly used in all the fields like, academics, Finance, Business, Hospitals, crime records, cyber security etc. [1]

Summarization is a technique to get concise and meaningful summary of any type of e-contents. There are two major approaches for summarizing text: Abstractive and extractive text summarization.[2]

TextRank is a graph-based extractive summarization algorithm. It is domain and language independent [3]. By implementing the TextRank algorithm user will get the shortened or summarized e-news with predefined ratio. This ratio will provide arbitrary length of output text.

II. LITERATURE STUDY

The study of text summarization started in 1958, which involved keywords, position of sentence, & word frequency. The online information is growing exponentially day by day, so it emerges with the problem of getting required information in minimum time. This text summarization may lead to help the user to get required information in less time and without changing its meaning. it seems difficult to summarize the text for Indian languages (low resource languages) due to limited availability of NLP tools and techniques for Indian languages.[4]

Pradeepika Verma, and Anshul Verma, described briefly about existing text summarization methods for Indian texts. They also shown the results of existing techniques of text summarization for Indian languages with English language and found that the NLP tools affects the performance of any summarizer. The result of precision, recall, and F1 measures for summarization methods for different languages is also discussed briefly.[4]

Sanjan S Malagi, Rachana Radhakrishnan, Monisha R, Keerthana S, proposed a system which covers a summarization model to produce an effective summary with the least redundancy and grammatically correct sentences from the source document. This approach has demonstrated good performance for most of the summarization purposes. This CMI model mainly demonstrates 4 modules: raw data, image data, news API and single text document.[5]

Mamatha Balipa, Dr. Balasubramani R, Harolin Vaz, Christina Shilpa Jathanna, attempted summarizing information from online health care forums about the disease Psoriasis to implement automatic text summarization. Online text is extracted using BeautifulSoup class available in urllib2 module. Then the topic of the text is confirmed to be Psoriasis by using Latent Dirichlet Allocation (LDA) algorithm.[8]

Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das and Apurba Sarkar, proposed a method which constructs a graph with sentences as the nodes and similarity between two sentences as the weight of the edge between them.[9]

Reda Elbarougy, Gamal Behery, Akram El Khatib, applied modified page rank algorithm with an initial score for each node that is the number of nouns in this sentence. More nouns in the sentence mean more information, so nouns count used here as initial rank for the sentence. Edges between sentences are the cosine similarity between the sentences, to get a final summary that contains sentences with more information and well connected with each other. [10]

Ahmed Elrefaiy, Ahmed Rafat Abas, Ibrahim Elhenawy, provided a review of collaborative survey which focuses on unsupervised techniques. It also describes evaluation of techniques of the summaries.[11]

Rasim Alguliev, Ramiz Aliguliyev, shown an approach which can improve the performance compared to sate-of-the-art summarization approaches. They have proposed new criterion functions for sentence clustering. They also have developed modified discrete differential evolution algorithm to optimize the objective functions.[12]

Kalliath Abdul Rasheed Issam, Shivam Patel, Subalalitha C. N., proposed technique which aims to capture all the varied information present in source documents. Also they have discovered that their model produces encouraging ROUGE results and summaries when compared to the other published extractive and abstractive text summarization models. [13]

Siddhant Upasani, Noorul Amin, Sahil Damania, Ayush Jadhav, A. M. Jagtap, obtained the rank or score of each sentence and the sentences with the rank above a particular value can be chosen to be included in the summary.[14]

Yash Asawa, Vignesh Balaji, Ishan Isaac Dey, surveyed numerous approaches, merits and limitations of the techniques of summarization. The Benchmark datasets of this domain and their features have also been examined. [15]

Rada Mihalcea and Paul Tarau, given 2 new methods for sentence and keywords extraction. And also shown that the results are favorably similar to other state of art defined algorithms. They have elaborated about the graphical representation of text and the how sentences are related with each other.[16]

Rada Mihalcea and Paul Tarau, used a graph-based ranking model for text processing, and shown how this model can be successfully used in natural language applications. They evaluated 2 advanced unsupervised methods for keyword and sentence extraction, and presented that the accuracy achieved by TextRank in these applications is competitive with that of previously proposed state-of-the-art algorithms.[17]

Chin-Yew Lin, stated the efficiency of applying sentence compression on an extraction based multi-document summarization system. The results shows that pure syntactic-based compression does not improve system performance.[18]

Deepali K. Gaikwad, Deepali Sawane and C. Namrata Mahender, seveloped a system for rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer. The paper shows technique which is used for generation of the appropriate question on given input/text.[19]

Yogeshwari V. Rathod [20] used sentence ranking algorithm to generate summary of Marathi news articles by extractive method. It gives effective summary in less time and with least redundancy.

Shraddha A. Narhari, Rajashree Shedge [21] proposed a text categorization of Marathi documents using LINGO & PCA algorithm. They proved this with improved results.

Jaydeep Jalindar Patil, Prof. Nagaraju Bogiri[22] used LINGO [Label Induction Grouping] algorithmfor improving results efficiently inmarathi text documents.

Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar [23] developed a system to Overcome the limitations of the lexical chain approach to generate a good summaryusing the WordNet thesaurus, pronoun resolution for news articles.

N. Dangre, A. Bodke, A. Date, S. Rungta, S.S. Pathak [24] proposed a System for Marathi News Clustering using Cluster algorithm to collect relevant Marathi news from multiple sources on web which results in enabling rich exploration of Marathi contents on web.

Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das and Apurba Sarkar, proposed a method which constructs a graph with sentences as the nodes and similarity between two sentences as the weight of the edge between them.[25]

Ahmed Elrefaiy, Ahmed Rafat Abas, Ibrahim Elhenawy, provided a review of collaborative survey which focuses on unsupervised techniques. It also describes evaluation of techniques of the summaries.[26]

Siddhant Upasani, Noorul Amin, Sahil Damania, Ayush Jadhav, A. M. Jagtap, obtained the rank or score of each sentence and the sentences with the rank above a particular value can be chosen to be included in the summary.[27]

Tomonori Kikuchi, Sadaoki Furui, Chiori Hori [28], proposed an automatic speech summarization method. It comprises of sentence compaction method by maximizing summarization score.

Sujata Lungare, Aman Jain, Tejashri Bhingare, Priyanka Tak, Mr.Pratik Kamble[29], given a brief view of how a user will be using a web application to summarize Marathi articles or documents using machine learning algorithm, where multimple documents are given as input and a meaningful summary is generated as an output.

III. PROPOSED SYSTEM

Automatic Text summarization (ATS) process is partitioned into 2 key parts. One is Language dependent and other is language independent. The foremost challenge with Indian languages summarizers is to do precise pre-processing and feature extraction. [4] For Marathi language, there are very limited essential tools and rich resources to summarize text.

Gensim = "Generate Similar" is a popular open source natural language processing (NLP) library used for unsupervised topic modelling. It uses top academic models and modern statistical machine learning to perform various complex tasks such as −

a) Building document or word vectors
b) Corpora
c) Performing topic identification
d) Performing document comparison (retrieving semantically similar documents)
e) Analysing plain-text documents for semantic structure.

Gensim depends on NumPy package for number crunching; also there are some core concepts that are needed to understand the Gensim library. That are document, corpus, vector and model. Here Corpus is collection of documents, vector shows, mathematical representation of document, and model refers to an algorithm used for transforming vectors from one representation to another. [16]

The document is an object of the text sequence type which is known as 'str' in Python. It can be a book, novel, paragraph, thesis, article etc.

The corpus is collection of documents, and its plural form is corpora. To represent every document of Corpus in a mathematical way user need vector representation. It will give the user features of input text so that it can be summarized. After vectorizing the Corpus the next step is to model it.

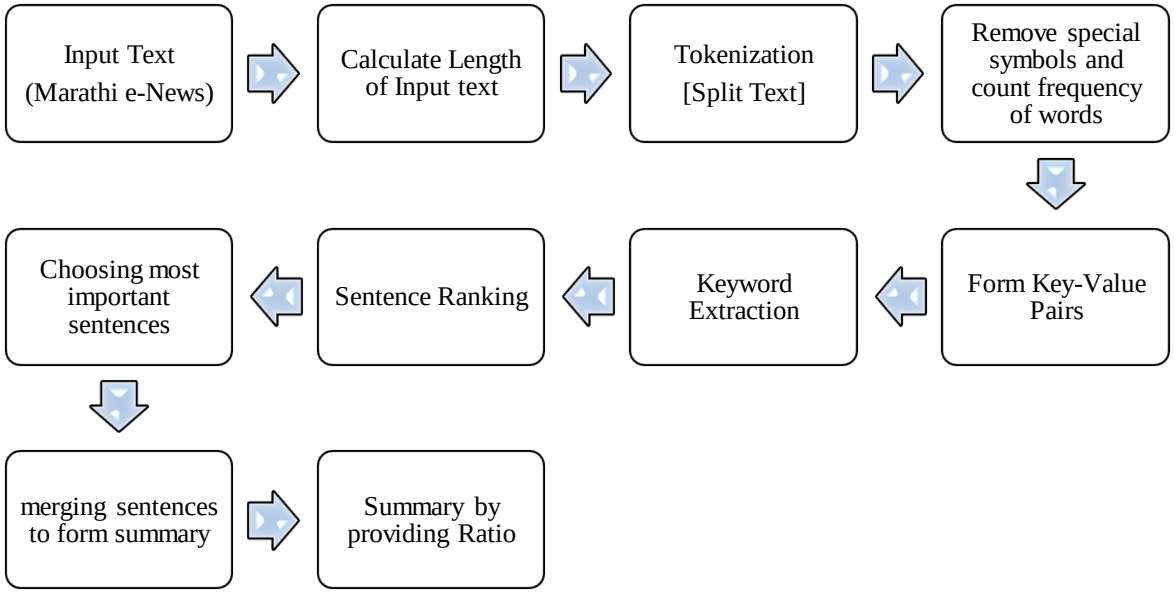

Fig1: Proposed system for Marathi e-news summarization

The proposed system shows the process of generating summaries of Marathi text. We are using dataset from github which has 1135 Marathi e-news articles. These are txt files and the contents are then passed on as an input for summarization. The figure above represents the proposed system for Marathi e-news summarization.

3.1 Experiments

For the task of automated summarization, TextRank models any document as a graph using sentences as nodes [6].

The experiments are done on the unstructured data to convert it into more comprehensible form that is structured data. The text rank algorithm follows the mentioned steps to get summary.

The first step is to input the e-news to be summarized, and calculate the length of the total input text. The input text length is compared with the summarized text so that the ratio can be verified.

The next step is to split the words by word_list=mytext.split(). It will separate each word in the sentence. The reason behind performing these steps is to pre-process the input data for applying the TextRank algorithm.

We Have to count frequency of each word because the irrelevant words i.e. An empty array is created for storing the count; to calculate this frequency count get () function is used and counter will help to get exact count of each word then.

Further, we make the Key-Value Pairs of the word and its frequency. Here, the efficiency of algorithm depends on the language user is using for text summarization. So, we are trying to improve the results by pre-processing data, and providing better efficacy of Marathi text.

3.2 Evaluation

There are numerous packages in the Python which are useful for this type of task. We are using some packages among them to summarize text. Primarily used packages are PyPI, Gensim, Corpora, Conda, Numpy, Scipy. The Numpy and SciPy are the 2 scientific computing Packages which must be installed before using Gensim package. The Corpora is also important package for summarizing text.

Gensim is a Python framework designed to help make the conversion of natural language texts to the Vector Space Model as simple and natural as possible. Gensim contains algorithms for unsupervised learning from raw, unstructured digital texts, such as Latent Semantic Analysis, Latent Dirichlet Allocation or Random Projections. These algorithms discover hidden (latent) corpus structure. [7] This framework gives user a better machine learning framework which enables efficient text summarization of Marathi. It helps to extract keywords by importing its framework.

A function to compute the similarity of sentences is needed to build edges in between. This function is used to weight the graph edges, the higher the similarity between sentences the more important the edge between them will be in the graph.[3]

IV. RESULT

This approach summarizes input text by using extractive summarization method. Here, there is a need of summarizing text by counting number of words, or by providing ratio which will be the value between 0 to 1.

If user uses the first method i.e. give count of words you require in summary, then the summary will be of the given count only.

In second method, if user is providing 0.1 value, it will give 10% summary of total input text. Likewise, if 0.8 value gives 80% of text in output. Ratio also plays important role in Gensim library.

The Extractive summarization is a best technique to summarize Marathi text as extracts important sentences and keywords and it do not change the meaning of input data. It will calculate the frequency of words, and filter out those high-ranking sentences for summary.

Input text:

mytext = केंद्रीय माध्यमिक शिक्षण मंडळ'तर्फे (सीबीएसई) इयत्ता दहावीचा निकाल आज, १५ जुलैला जाहीर केला जाणार आहे. निकालाची नेमकी वेळ मात्र बोर्डाने जाहीर केलेली नाही. बारावीचा निकाल सोमवारी, १३ जुलैला जाहीर झाला होता.

दहावीचा निकाल 'सीबीएसई'चे अधिकृत संकेतस्थळ CBSE.NIC.IN येथे दिसेल. यासोबतच निकालासाठी स्वतंत्र पेज असलेल्या CBSERESULTS.NIC.IN या लिंकवरही तो पाहता येईल. यंदा 'सीबीएसई'ने आयव्हीआरएस सुविधा उपलब्ध करून दिली आहे. यासाठी विद्यार्थ्यांना ०११-२४३००६९९, ०११-२८१२७०३० या क्रमांकांवर संपर्क साधावा लागेल.

कॉल सुरू असतानाच विचारले गेल्यावर मोबाइलवर आपला रोल नंबर आणि जन्मतारीख टाकल्यानंतर निकाल समजेल.

The Extractive text summarization by providing ratio is as follows:

केंद्रीय माध्यमिक शिक्षण मंडळ'तर्फे (सीबीएसई) इयत्ता दहावीचा निकाल आज, १५ जुलैला जाहीर केला जाणार आहे.निकालाची नेमकी वेळ मात्र बोर्डाने जाहीर केलेली नाही. बारावीचा निकाल सोमवारी, १३ जुलैला जाहीर झाला होता.दहावीचा निकाल 'सीबीएसई'चे अधिकृत संकेतस्थळ cbse.nic.in येथे दिसेल.यासोबतच निकालासाठी स्वतंत्र पेज असलेल्या cbseresults.nic.in या लिंकवरही तो पाहता येईल.यंदा 'सीबीएसई'ने आयव्हीआरएस सुविधा उपलब्ध करून दिली आहे.

The following table shows the comparison between lengths of input text and output text by using Ratio:
Ratio = 0.8

Table1: Comparison of length of input text and summarized text using Ratio.

Length of Input text	Length of Output text by using Ratio
607	418

Graphical representation of the table1 is shown below:

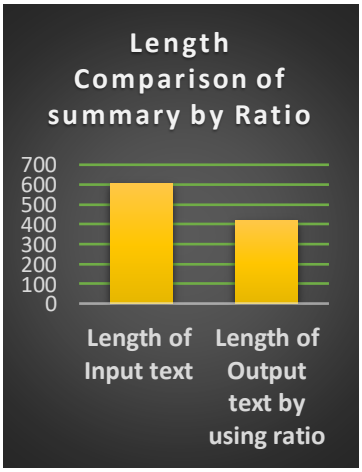

Fig2: Comparison of length of input text and summarized text using Ratio

The above illustrations clearly show the efficient Extractive summarization system for Marathi text using TextRank algorithm using Gensim library. If the ratio value change, the result will also have influence of this value.

V. CONCLUSION

In this paper we described the Automatic text summarization of Marathi e-news articles using TextRank algorithm with emphasis on the Gensim library. The proposed system evidences the sully unsupervised extractive text summarization by using the method which provides the ratio value which in between 0 and 1, to get summary of input text. Here we followed the pre-processing and processing of input Marathi text. Pre-processing includes calculating length, removing special symbols, and tokenization of input text. Processing defines finite steps like forming key-value pairs with corpus and modelling it. The sentences are then ranked and according to the ranking, the important sentences are merged to get summary. This is a very precise method which is specially used for Marathi text, as the techniques and tools perform in a different way with different languages.

REFERENCES

[1] Apurva D. Dhawale, Sonali B. Kulkarni, and Vaishali Kumbhakarna, "Survey of Progressive Era of Text Summarization for Indian and Foreign Languages Using Natural Language Processing", ICIDCA 2019, LNDECT 46, pp. 654–662, Springer Nature Switzerland AG, 2020.

[2] Apurva D. Dhawale, Sonali B. Kulkarni, and Vaishali Kumbhakarna, "A Survey of Distinctive prominence of Automatic Text Summarization Techniques Using Natural Language Processing", International conference on mobile computing and sustainable informatics, (ICMCSI), Springer conference, Tribhuvan university, Nepal, 2020.

[3] Federico Barrios, Federico L_opez, Luis Argerich1, Rosita Wachenchauzer,"Variations of the Similarity Function of TextRank for Automated Summarization", ASAI 2015- 16 Simposio Argentino de Inteligencia Artificial, February 2016.

[4] Pradeepika Verma, and 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.

[5] Sanjan S Malagi, Rachana Radhakrishnan, Monisha R, Keerthana S, "Content Modelling Intelligence System Based on Automatic Text Summarization", Int. J. Advanced Networking and Applications Volume: 11 Issue: 06 Pages: 4458-4467 ISSN: 0975-0290, 2020.

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

[7] Https://Pypi.Org/Project/Gensim/0.6.0/

[8] Mamatha Balipa, Dr. Balasubramani R, Harolin Vaz, Christina Shilpa Jathanna, "Text Summarization For Psoriasis Of Text Extracted From Online Health Forums Using Textrank Algorithm", International Journal Of Engineering & Technology, 7 (3.34) (2018) 872-873, 18 September 2018.

[9] Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das And Apurba Sarkar, "Graph-Based Text Summarization Using Modified Textrank", J. Nayak Et Al. (Eds.), Soft Computing In Data Analytics, Advances In Intelligent Systems And Computing 758, Springer Nature Singapore Pte Ltd. 2019.

[10] Reda Elbarougy, Gamal Behery, Akram El Khatib, "Extractive Arabic Text Summarization Using Modified Pagerank Algorithm", Egyptian Informatics Journal 21, 73–81, Science Direct, Elsevier, (2020).

[11] Ahmed Elrefaiy, Ahmed Rafat Abas, Ibrahim Elhenawy, "Review Of Recent Techniques For Extractive Text Summarization", Journal Of Theoretical And Applied Information Technology 15th December 2018. Vol.96. No 23, Issn: 1992-8645, Jatit & Lls, 2005.

[12] Rasim Alguliev, Ramiz Aliguliyev, "Evolutionary Algorithm for Extractive Text Summarization", Intelligent Information Management, 1, 128-138, Scientific Research, SciRes, 2009.

[13] Kalliath Abdul Rasheed Issam, Shivam Patel, Subalalitha C. N., "Topic Modeling Based Extractive Text Summarization", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-6, April 2020.

[14] Siddhant Upasani, Noorul Amin, Sahil Damania, Ayush Jadhav, A. M. Jagtap, "Automatic Summary Generation using TextRank based Extractive Text Summarization Technique", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 May 2020.

[15] Yash Asawa, Vignesh Balaji, Ishan Isaac Dey, "Modern Multi-Document Text Summarization Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020.

[16] https://www.tutorialspoint.com/gensim/gensim_quick_guide.htm

[17] Rada Mihalcea and Paul Tarau, "TextRank: Bringing Order into Texts", Anthology ID: W04-3252; Volume: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain, 404–411, July 2004.

[18] Chin-Yew Lin, "Improving Summarization Performance by Sentence Compression –A Pilot Study", University of Southern California/Information Sciences Institute, 2003.

[19] Deepali K. Gaikwad, Deepali Sawane and C. Namrata Mahender, "Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 51-54, 2018.

[20] Yogeshwari V. Rathod, "Extractive Text Summarization of Marathi News Articles", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 07,July 2018.

[21] Shraddha A. Narhari, RajashreeShedge, "Text Categorization of Marathi Documents using Modified LINGO", IEEE, 2017

[22] Jaydeep Jalindar Patil, Prof. NagarajuBogiri, "Automatic Text Categorization-Marathi documents", International Conference on Energy Systems and Applications (ICESA 2015), IEEE, 2015.

[23] Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar, "Automatic Text Summarization of News Articles", International Conference on Big Data, IoT and Data Science (BID) Vishwakarma Institute of Technology, Pune, Dec 20-22, IEEE, 2017

[24] N. Dangre, A. Bodke, A. Date, S. Rungta, S.S. Pathak, "System for Marathi news clustering", 2nd International conference on Intelligent computing, communication & convergence, bhubaneshwar, ELSEVIER, 2016.

[25] Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das And Apurba Sarkar, "Graph-Based Text Summarization Using Modified Textrank", J. Nayak Et Al. (Eds.), Soft Computing In Data Analytics, Advances In Intelligent Systems And Computing 758, Springer Nature Singapore Pte Ltd. 2019.

[26] Ahmed Elrefaiy, Ahmed Rafat Abas, Ibrahim Elhenawy, "Review Of Recent Techniques For Extractive Text Summarization", Journal Of Theoretical And Applied Information Technology 15th December 2018. Vol.96. No 23, Issn: 1992-8645, Jatit & Lls, 2005.

[27] Siddhant Upasani, Noorul Amin, Sahil Damania, Ayush Jadhav, A. M. Jagtap, "Automatic Summary Generation using TextRank based Extractive Text Summarization Technique", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 07 Issue: 05 May 2020.

[28] Tomonori Kikuchi, Sadaoki Furui, Chiori Hori, "Two-Stage Automatic Speech Summarization By Sentence Extraction And Compaction", Research gate, March 2014.

[29] Sujata Lungare, Aman Jain, Tejashri Bhingare, Priyanka Tak, Mr.Pratik Kamble, "Multi-Document Text Summarization of Marathi Regional Language", International Journal Of Innovative Research In Technology, Ijirt, Volume 5 Issue 12, Issn: 2349-6002, May 2019.