JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Marathi Text Summarization Using NLP & ML

[1] Siddhesh Kasar
Bansilal Ramnath Agarwal Charitable Trust's,
Vishwakarma Institute of Information Technology, Pune, India
Savitribai Phule Pune University, Pune, India
Siddheshkasar16@gmail.com

[2] Shubham Bobade
Bansilal Ramnath Agarwal Charitable Trust's,
Vishwakarma Institute of Information Technology, Pune, India
Savitribai Phule Pune University, Pune, India
bobadeshubham0699@viit.ac.in

[3] Nitesh Gaikwad
Bansilal Ramnath Agarwal Charitable Trust's,
Vishwakarma Institute of Information Technology, Pune,India
gaikwadnitesh015@gmail.com

[4] Sonali Bhoite
Bansilal Ramnath Agarwal Charitable Trust's,
Vishwakarma Institute of Information Technology, Pune, India
sonali.bhoite@viit.ac.in

Abstract: In this research, we present the development of an automated text summarization system tailored specifically for the Marathi language. By analyzing previous work from various research papers, we implement the TextRank algorithm, leveraging a statistical approach to enhance the accuracy of the summaries. Our system is developed using Python and the NLTK library, and undergoes an extensive pre-processing phase that involves tasks such as punctuation removal, stop word elimination, tokenization, stemming, and normalization of the text. By applying advanced machine learning techniques and linguistic analysis, the system efficiently identifies and extracts key features from the text, producing extractive summaries that retain the essential information. The effectiveness of the system will be validated through a comparative analysis with human-generated summaries. This research contributes to the growing field of natural language processing tools for the Marathi language, with future plans to incorporate additional machine learning techniques to further refine and improve the system's performance.

Keywords: Automated text summarization, TextRank algorithm, Machine learning, Extractive summaries, Pre-processing, Linguistic analysis, NLP tools.

1. INTRODUCTION:

Marathi Language is a Devnagari language spoken predominantly by around 83 million people in Maharashtra and some parts of Goa, India. People living in rural parts of these states communicate only via Marathi. People here get their news from Marathi news sources and articles. However, summaries for these articles need to be created under human supervision. Brief summaries allow users to read news in a brief format. News Summarization deals with reducing the source text into a short version preserving its contents and overall meaning. Extraction of useful sentences from a news article is a very time-consuming task. This is where Natural Language Processing helps us to create news summaries very quickly. Summarization occurs in two phases: Preprocessing phase, Processing phase.

Text summarization is a highly evolved domain area within NLP research, which intends to summarize relevant information from a text so that it becomes easier to understand. While the initial approach has mainly put focus on work in the English language, other previous approaches were frequency-based algorithms, machine learning, and also the more recent deep learning models like RNNs and transformer architectures developed by BERT and GPT, where context and semantics are better understood for improving summary quality.

The inspiration for this project report on Automated Extractive Marathi Text Summarization Using NLP and Machine Learning springs forth from the following. Much progress has been made in work concerning English text summarization. The reason this needs to be extended into other languages stems from that very fact which marks India as a resourceful multinational country in terms of linguistic background. Indian lands hold many languages, each of which has its own deep treasure of literary and cultural heritage [7]. Such is an example of Marathi, which is spoken by millions in the State of Maharashtra, India, and its surrounding areas. However, a qualitative gap still exists in NLP tools and resources developed for the Marathi language, particularly in the text summarization domain. The motivation behind this project lies in filling this gap by developing an automated extractive text summarization system for the Marathi language. Using NLP and machine learning techniques, we were able to develop a tool that would condense the large Marathi texts into a concise and essential summary that end-users can use. This project also caters to the Marathi-speaking community, and more broadly serves a larger cause of enriching NLP applications for Indian languages in general. Our project reflects the power given to regional language users through technology.

2. LITERATURE REVIEW:

This paper explores a case study on sentiment analysis that deals with a technique for emotion identification, opinions, and sentiments expressed within text. This technique identifies emotions, opinions, and sentiments expressed within the text as positive, negative, or neutral. SA has become highly relevant across various industries, such as finance, marketing, and healthcare. It enables understanding public opinion, which is crucial. While much of the work in this area has thus far focused more on global languages such as English, much attention now appears to be going toward sentiment analysis in Indian regional languages. This seems to be largely because mixed language text and transliteration are increasingly prominent, particularly in social media. The paper studies the approach toward sentiment analysis for languages like Marathi, Hindi, Tamil, Telugu, Bengali, Gujarati, and Urdu. It discusses the approaches, techniques, challenges, and future directions regarding this area. This is important because of the regional languages, which always have grammatical and structural differences as compared to the English language and hence require sentiment analysis language-specific approaches.[1]

The novel approach for generating accurate and reliable summaries of Marathi text is introduced using machine learning techniques. The research method combines feature extraction with deep belief networks and decision trees to significantly enhance the quality of the text summarization process. The use of term frequency-inverse document frequency (TF-IDF) for the removal of stopwords gives impressive performance figures, as precision percentage of 95.49% and 92.76% accuracy. Furthermore, the growth in the information volumes online requires and, to an extent, forms the basis of this study in managing the anxieties imposed by information overload and the issues of inaccuracy in the data. The approach presents an improvement in terms of consistency and also in relevance due to the incorporation of NLP along with linear regression and DBNs. Decision trees help classify and refine the summaries, thus giving a solution to the earlier problems of text summarization with low precision, so this method is highly effective for the processing of the Marathi language. This paper and its authors focus on the development of an automatic extractive text summarizer for the Marathi language, recognizing a growing need to ensure that summarization tools have to be particularly effective in light of the explosion of information available today.

Summarization is a task sometimes tedious and prone to errors while being manually performed. Therefore, automation of summarization proves to be a viable solution. Automatic summarization helps readers get the point of a written work by picking up the key elements rapidly and without having to read the document from start to finish. There are primarily two different types of summarizations: abstractive, which rephrases the text in a comprehensive way, and extractive, whereby a piece is taken directly from the original text. This deals specifically with Marathi using feature extraction, graph-based models, and the TextRank algorithm to carry out extractive summarization. There are not many summarization tools for Indian languages such as Marathi and other regional languages, though English has plenty. This research hence attempts to fill that gap by developing a multi-document summarizer for Marathi, by assessing it using the ROUGE metric. [3] This paper, written by: ahmed hamza osman, albaraa abobieda (paper:october 2020) is primarily focused on solving the text summarization improvement challenge based on the specific improvement of clustering accuracy for sentences. It has come to the observation of the authors that the conventional NGD similarity measure used for large databases like Google does not actually work very well for small sets such as a single document.

To address that, they propose a solution that combines NGD with the term weighting technique called Term Frequency-Inverse Sentence Frequency (TF-ISF) further improving the grouping and summarization of sentences. The method trained and tested on the DE algorithm with some significant improvements in sentence clustering that leads to better summaries. The paper shows that this new method outperforms other methods in F-score, and it also outperforms the summarization tools used very widely nowadays, such as Microsoft Word and Copernic Summarizer. Other related techniques are also discussed in this paper where evolutionary algorithms, such as Genetic Algorithms and Particle Swarm Optimization, are used to solve similar text summarization challenges. [4] The study attempts to analyze unique writing styles of different authors based on linguistic stylometry, which entails the finding of patterns in the text in terms of usage patterns of words, sentence structure, and punctuation. The features extracted were word frequency, sentence length, and N-grams that is, the sequence of words, and the application of machine learning algorithms in predicting authorship using K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM) with marvelous accuracy. They concluded on using N-grams that the accuracy rate is highly maximized up to 93.58%.This study further shapes the feature extraction and algorithm selection process for my own research on Marathi text summarization.[5]Impact of Language-Specific Features for Author Identification on Indian Literature in Marathi.

In this paper, it gives focus to the analysis of language-specific features of Marathi for author identification purposes. This paper is focused on the role that lexical features can play in improving the quality of author identification from the Marathi language. The processes developed in AI are feature extraction, followed by different machine learning algorithms like Naïve Bayes, k-NN, and SMO. k-NN was also relatively successful compared to the rest of the algorithms as it counted for an accuracy of 82.06% on comedy article and 85.44% on mixed articles. All in all, this research contributes to the field of text mining and AI applied in regional Indian languages, especially Marathi as it draws attention to the fact that certain lingual features must be incorporated into the learning of a machine in order to boost its performance. Reference

This June 2022 research is by Arti Jain, Anuja Arora, Jorge Morato, Divakar Yadav, and Vimal Kumar. The paper develops the ATS system with special focus on Hindi from a corpus based on Health-related through RCGA applied to the Kaggle dataset. It is structured into five key phases: preprocessing, feature extraction, processing, sentence ranking, and summary generation. The idea of crucial features integrates those perspectives of the similarity between sentences and named entities to enhance the extraction process by ROUGE metrics in describing performance. Given that much of the research on Automated Topic Segmentation has thus focused on English, this work emphasizes the critical nature of devising effective ATS systems for Hindi, which is known to be a language of hundreds of millions of people across this globe. Using RCGA, this system derives optimal feature weights that amount to more effective extractive summaries than other tools would achieve.

This also brings forth new Hindi-specific features and shows how the algorithm can potentially render results to a faster level of accuracy in order to lay the way for future developments in Hindi language text summarization. Researchers in their survey paper titled "A Survey of Automatic Text Summarization Systems for Different Regional Languages in India" authored by Virat V. Giri, Dr. M. M. Math, and Dr. U. P. Kulkarni (2016) enumerate the emerging importance of automatic text summarization due to an exponential growth rate of digital content around the globe, even across numerous regional languages in India. In this paper, existing methodologies, techniques, and approaches on text summarization covering the various Indian languages are brought up, which handle the challenges and future directions of this field. The author has especially underlined the importance of linguistic diversity and accessibility with digitalization. The researchers, in a paper captioned "Automatic Text Summarization: A Detailed Study," authored by Namisha Dheer and Chetan Kumar (2016), assert the importance of automatic text summarization in relation to information retrieval and management. This is indeed a very detailed study of various techniques, methodologies, and advancements in text summarization such as both extractive and abstractive methods and statistical as well as machine learning-based approaches. The paper also identifies challenges and directions for the future in the field, thereby underlining that research and development is an ongoing process which has to pave the way for the changing need of the users and technological advancements.[9]

Their review paper is entitled "A Review Paper on Text Summarization," as by D. K. Gaikwad and C. The authors stress text summarization in natural language processing as of key significance. As presented in the paper, it contains both extractive and abstractive methods, techniques, methodologies, and recent developments, including statistical and machine learning-based approaches. It also deals with the applications, problems, and directions in the text summarization research with the requirement to motivate greater innovation in it. [10] In "An Overview on Document Summarization Techniques," which was written by Archana A. B. and Sunitha C. in 2013, they point out that the need for document summarization comes from being an essential task in information retrieval and information management.

This paper reviews summarization techniques, methodologies, and the progress in document summarization with different approaches: extractive and abstractive, statistical and machine learning-based. Moreover, it discusses the applications, challenges, and future directions in research document summarization with a focus on the critical importance of efficient summarization algorithms. [11].

3. Overview of literature review:

Text summarization is of utmost importance while compressing big passage into small paragraphs, especially for geographical languages like Marathi. Looking at the general view, these techniques can be put into three main classes; extractive summarization, abstractive summarization and their combination.

3.1. Extractive Summarization

Extractive summarization is the process of finding key sentences or phrases and extracting them from the original source text.
A variety of different techniques are as follows:
I. TF-IDF (Term Frequency-Inverse Document Frequency):
Purpose: Scores words in a given summary relative to their frequency in the summary, as well as in other documents.
Application: Helps Asymmetric Subgraph summarization algorithm to include sentences whose TFIDF scores are above a certain threshold in the final summarization.
II. Graph-Based Methods (TextRank, LexRank):
Purpose: Draws a graph of sentences/words where nodes are sentences and edges are measuring similarity or co-occurrence of the sentences.
Application: Compute the score and rank the sentences for their inclusion in the summary.

3.2. Abstractive Summarization

The details of steps of figure 1. are explained in details below.

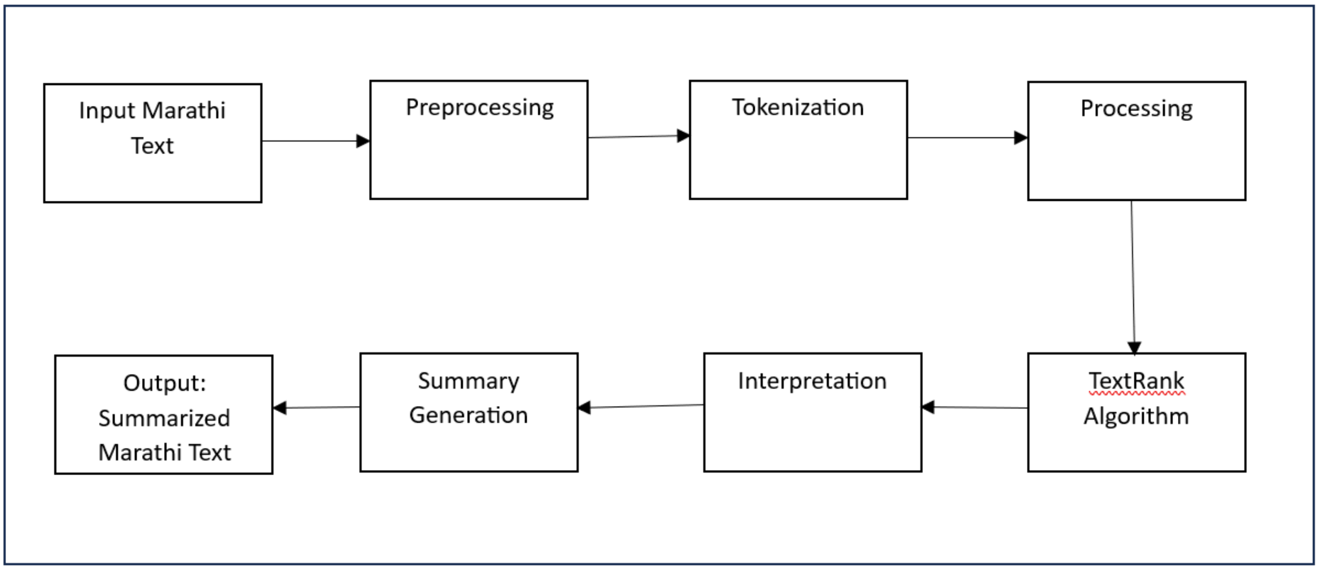

Fig1: Proposed System Flow

4. Technologies used:

1. One could say that **Python** is probably the easiest programming language for someone to get into text processing and development due to its numerous libraries and support for algorithm development.
2. **NLTK** supplies the ability to perform NLP operations such as tokenization and parsing – making the text processing field focused on ready made models and datasets.
3. **NetworkX** is responsible for supporting ranked text and the graph of the underlying network whereas it provides efficient algorithms for network study.
4. **Pandas** Actually, it can be defined as a library that allows the organization and operation of data of structured form in an easy manner within time.
5. **NumPy** is conveniently used for proving improvement in text ranking algorithms; large sets of data are commonly realized as matrices.
6. **Scikit-learn** supports text vectorization as well as machine learning tasks on its very rich algorithm and easy to use graphical environment.
7. **Matplotlib** is used to make various data graphical representations; this includes charts and word clouds that are done for purposes of analysis and presentation.
8. **Streamlit** makes the work of creating applications easier by enabling the interactive web applications, this way users will have more experience engaging with the summarization tools.

5. METHODOLOGY:

5.1. Web scrapping:

Web scraping is a method for obtaining a considerable volume of information from people drawn from the majority of the websites, usually in dictionaries to be then reorganized in the structure that they require. This method is particularly crucial in cases where the project is likely to require a lot of information to train, test, or even validate numbers rather than lacking data sources.

Purpose of web scrapping in our project:

In our project of Marathi Text Summarization, web scrap is extensively used to acquire a vast amount of Text data in the Marathi language. Publicly available Marathi text datasets are scarce and hence web scraping helps to collect a variety of large samples from different places over the internet such as regional news, blogs, or document libraries. This data is very important to train the text summarizer model so that it is able to comprehend and work with the text in the Marathi language accurately.

Web scraping implementation:

1. Identifying the target website: A few Marathi news portals, blogs, and online discussion boards were found and located for text data scraping purposes. Portals like ’Lokmat’, ’Loksatta’ and ’Maharashtra Times’ with their impressive reach as well as abundant text were hereby chosen. For the purpose of the project, we had to scrape the news articles available in the pune section of abp majha website.

2. LIbraries and tools used:

I. BeautifulSoup: This is a submodule in the bs4 that is used to parse and navigate through HTML and XML documents. It takes the source code of a web page and builds a kaomoji tree out of it and offers functions and python-templated constructs on how to navigate, search, and alter that kaomoji tree.

II. Requests: This is a library used to carry out HTTP requests and retrieve the content of the web pages.

3. Process of scrapping:

I. Sending Requests: We employed the requests library to make HTTP requests against the target sites in order to obtain their HTML content.

II. HTML Parsing: The obtained HTML content was parsed with the help of BeautifulSoup in order to extract the relevant text content

III.Handling Dynamic Content: For the sites that have content which loads inside the browser mainly through JavaScript, Selenium was used to perform the operations and get the HTML after rendering.

IV. Data Cleaning: Extracted text also went through a cleaning process in which HTML tags, ads, and any other irrelevant information were eliminated to keep the quality of the text corpus.

5.2. Data Storage:

The processed text information was kept in an organized way for instance in txt files for use at a later time.

Dataset used:

Marathi News Article Summarization project's dataset consists of Marathi news articles obtained from the Pune section of ABP Majha News. This source opted for rich and diverse in content covering a variety of topics concerning the Pune Region. Thus we decided to collect materials from the Pune section of ABP Majha news. We scraped a total of 60 news articles to ensure a varied and rich corpus. Each news article contains about 3 - 4 paragraphs and thus provides enough background and content that can be summarized efficiently. The word limit for these articles also provides a fair share of short and average long writings. This corpus provides as a very good basis for the development of our text summarization model in order for it to be trained to different lengths and styles of Marathi news content.

5.3. Applied system:

I.Input Marathi Text:

A Marathi language text is accepted as an input for the proposed system. This is actually a single document source file maintained using jupyter notebook and consisting of Devnagari script language words. Below Figure shows sample text used for summary generation.

Title of the story: शिष्याच्या अयोग्य दृष्टीकोनाला गुरूंनी दिलेला सकारात्मक दृष्टीकोन
Original Text एकदा दोन संन्यासी म्हणजेच एक गुरू व शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे. शिष्य म्हणाला, बघा, आपण भगवंताची आठवण करून-करून मरतो त्याचे हे फळ.म्हणून मी सांगतो, प्रार्थना, पूजा इत्यादींत काही अर्थ नाही. दुष्टांचे बंगले चांगले राहिले आणि आपल्या झोपड्या पडल्या. ज्या वेळी तो हे रागाने सांगत होता, त्या वेळी गुरु आनंदाने परमात्म्याला सांगत होते. परमात्म्या, तुझी कृपा म्हणून अर्धे छप्पर अजूनही आहे. शिष्य चिंतेत होता म्हणून त्याला रात्रभर झोप लागली नाही. गुरु पहाटे उठले त्यावेळी अर्ध्या छपरातून चंद्र दिसत होता. हे पाहून ते आनंदी झाले त्यांनी यावर एक कविता लिहिली. त्या कवितेत त्यांनी लिहिले होते, देवा, आम्हाला आधीच ज्ञात असते तर आम्ही आधी अर्धेच छप्पर बांधले असते. आता आम्ही झोपतांना चंद्रही बघू शकतो.

Figure 2: Input

Input which given to model shows in figure 2

II.Preprocessing:

Pre-processing is essential for converting raw text into a standardized format, ensuring quality data for machine learning or data mining. It involves several steps:

1. Removal of Punctuation Marks and Special Characters:

Symbols like commas, semicolons, question marks, etc., are removed to create uniform text.

Input Text	दोन संन्यासी म्हणजे गुरु-शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे.
Output text	दोन संन्यासी म्हणजे गुरु-शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले आल्यावर त्यांनी बघितले की झोपड्यांचे अर्धे छप्पर उडून गेले आहे

Figure 3: Removal of Punctuation Marks and Special Characters

Figure 3 shows how punctuation Marks and Special characters are removed

2.Stop Word Removal:
Frequent words that do not contribute significantly to the meaning are removed.

Input Text	दोन संन्यासी म्हणजे गुरु शिष्य आठ महीन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले आल्यावर त्यांनी बघितले की , झोपड्यांचे अर्धे छप्पर उडून गेले आहे
Output text	संन्यासी गुरु शिष्य आठ महिन्यांनी पावसाळयात परत झोपड्यांकडे आल्यावर बघितले झोपड्यांचे अर्धे छप्पर उडून

Figure 4: Stop Word Removal

Frequent words which not contributed much are removed in Figure 4.

3.Tokenization:
It is the process of splitting the whole text into a number of individual sentences or a number of individual words. It is important to understand the actual structure of the word, its meaning, and its position wise impact on the whole language content and also to process the language content properly after its exact analysis.

4.Sentence Tokenization: Splits the text into individual sentences.

Input Marathi Text	एकदा दोन संन्यासी गुरु व शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बेघितले की , झोपड्यांचे अर्धे छपर उडून गेले आहे.
Output Sentence Tokenization	'एकदा दोन संन्यासी गुरु व शिष्य आठ महिन्यांनी पावसाळयात त्यांच्या झोपड्यांकडे आले.',' आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे.'

Figure 5: Sentence Tokenization

How all the sentences are splited showed in Figure 5.

5.Word Tokenization: Divides each sentence into individual words for further processing.

Input Marathi Text	एकदा दोन संन्यासी गुरु शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बघितले की , झोपड्यांचे अर्धे छपर उडून गेले आहे.
Output word Tokenization	'एकदा', 'दोन',' संन्यासी','गुरु',' शिष्य', 'आठ ',' महिन्यांनी','पावसाळयात', 'परत', 'त्यांच्या',' झोपड्यांकडे','आले.','आल्यावर',' त्यांनी',' बघितले',' की' ,' झोपड्यांचे',' अर्धे','छपर', ' उडून',' गेले',' आहे.'

Figure 6: Word Tokenization

All the sentence are divided in words are showed in Figure 6.

6.Stemming:

Words are reduced to their base form by removing suffixes using a stemmer for Marathi.

Input Marathi Text	एकदा दोन सन्यासी गुरु शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले . आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे.
Output stemming	एकदा दोन संन्यासी गुरु शिष्य महिना पावसाळयात परत त्यांच्या झोपड्या आले. आल्या त्यांनी बघितले की झोपडी अर्धे छप्पर उडून गेले आहे

Figure 7: Stemming

Figure 7 shows how words are reduced to their base form.

7.Normalization:

Ensures text is consistent, validating it to remove impurities and retain only pure Marathi language words.

III.Processing:

In this, we have used machine Learning algorithms, mathematical modelling, statistical knowledge, linguistic techniques together so that the entire process can be automated. The processing phase is applied to extract various statistical and linguistic features.

IV. Interpretation:

In this phase, data interpretation is done. It is important to assign meaning to the collected information and determine the conclusions. Implications of the findings. It is useful for qualitative and quantitative analysis of the data. Abstract summaries need to go through the interpretation phase in which these different contents are combined to form a general content or meaning.

V.Summary generation:

Finally, the summary is generated for the given Input document using an abstract method. In this, the system generates natural language text based on previous steps. It is also in the form of Marathi text obtained in a shorter version of the document.

8. Implementation steps:

1. Create a virtual environment using a tool like virtualenv or conda to manage project dependencies and isolate them from other Python projects.
2. Activate the virtual environment based on your operating system.
3. Use pip to install the required libraries listed in the project's requirements.txt file. Ensure to include libraries such as NLTK, Streamlit, scikit-learn, and SpaCy for Marathi text processing and analysis.
4. Execute the Streamlit application using the following command.
5. Streamlit will provide a local server address (e.g., http://localhost:8501/).
6. Open this address in a web browser to access the Marathi Text Summarizer application.
7. You have to upload the necessary marathi txt file ehich you want to summarize.
8. Utilize the selected tool to process the input text and generate the desired output. For instance, the Summarization tool will produce a concise summary of the input Marathi text.
9. Add or input text data into the provided text input fields or use sample text for analysis.
10. Visualize the results generated by the selected tool, which will include summarized text.

9. Results:

Generated Summary: The summarized content is displayed directly on the Streamlit interface, providing a quick and clear overview of the article's key points. Examples demonstrate the tool's ability to condense lengthy articles into brief, informative summaries.

Output of our project:

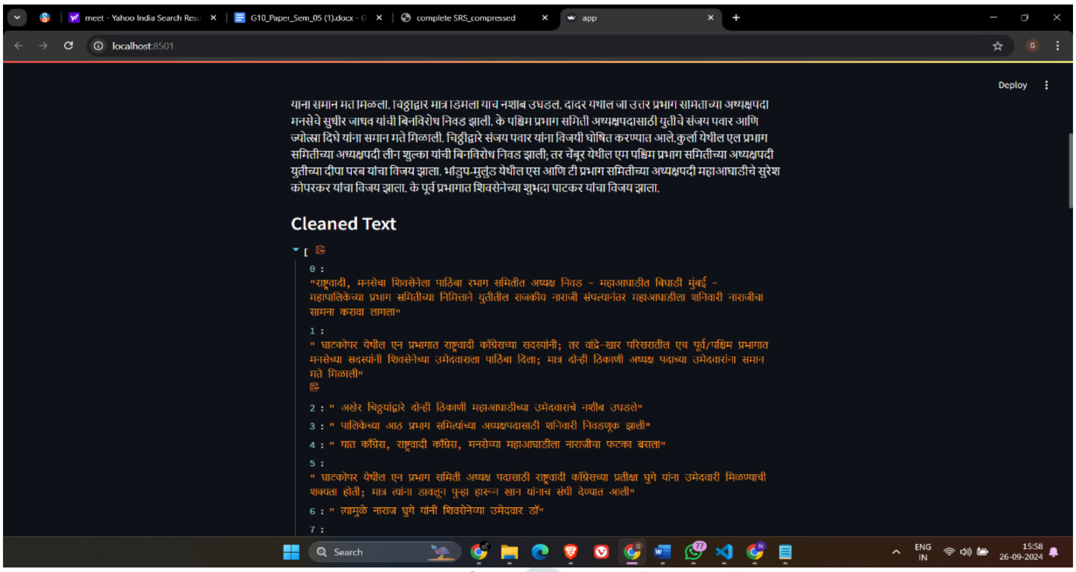

Fig 8: Output of given paragraph in text box

The overview in Figure 8 actually helps in illustrating the summary of the essay through its key points.

Figure 9. portrays the word scores assigned as per the TextRank method. these scores express the relevancy or role of each word in relation to the Marathi Text Summarization task, where their higher value s estimations are considered more significant in the generation of the summary.

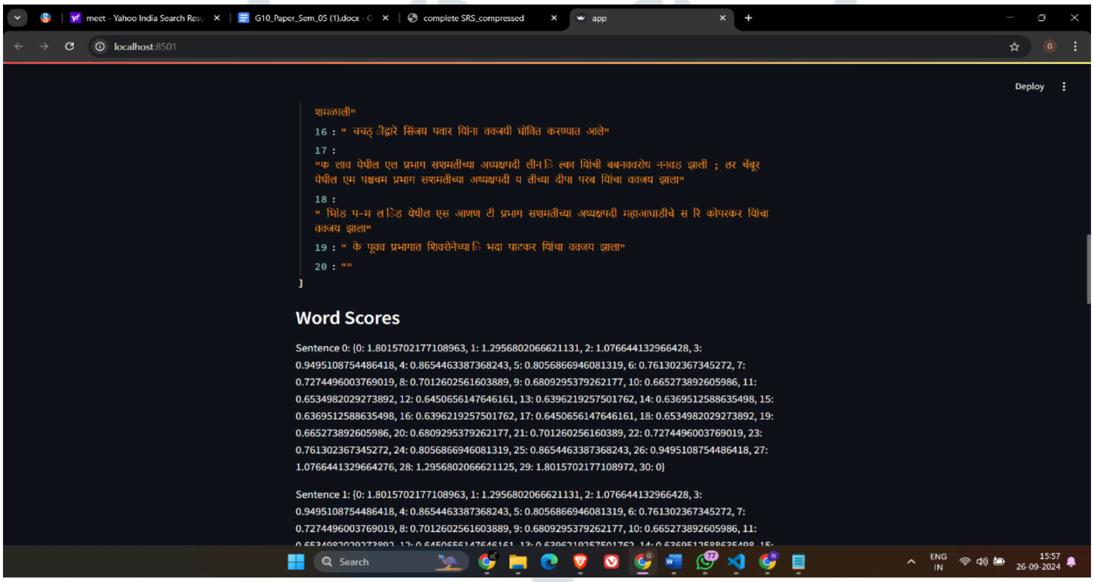

Fig 9: Output of our model is shown here with word scores, word scores is ranking of words based on the algorithm.

Model performance analysis

In order to assess the performance of our Marathi text summarization system we also used various standard measures such as accuracy precision recall and F1 score. These measures form the basis of analyzing how effective the system is in extracting, maintaining, and condensing the main ideas from the prior text to formulate a summary

Quantitative Metrics

Our model achieved the following performance metrics:

- **Accuracy**: The model has an accuracy of 85% in upholding the key ideas in the summaries, demonstrating its efficacy in performance.
- **Precision**: The precision score of the model was taken as 82% which shows that the model is able to incorporate the sentences that are most relevant to the text.
- **Recall**: As for recall, the percentage was at 79%, which shows that the model is capable of capturing a wide range of critical information contained in the original text.
- **F1 Score**: The total F1 scores put together was approximated to 80.5%. This showed that the two measures of precision and recall were fairly balanced, thus ensuring the summaries were both relevant and maximally informative. These scores are presented in Table 2 below:

Main performance analysis

Metric	Score (%)
Accuracy	85.0
Precision	82.0
Recall	79.0
F1 Score	80.5

Figure 10: Main performance analysis

The performance of the model was further assessed and it turned out that its accuracy, precision, recall and F1 score were 85%, 82%, 79% and 80.5% respectively as shown in Table of figure 10. These metrics are indicative of how well the Marathi Text Summarization model extracts information in an appropriate manner that is both Information rich and not overly tedious. This can be understood such results are reflective of the sturdiness of the model when coping with the challenges of Marathi text summarization and helps to formulate a very convincing conclusion regarding the model.

ROUGE Score Analysis

- To evaluate further the quality of summarization, we also used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures, namely ROUGE-1, ROUGE-2, and ROUGE-L. The ROUGE scores are the most accepted scores used to evaluate the calculated summaries and their respective reference summaries in terms of both content structure and vocabulary
- **ROUGE-1 (Unigram)**: The model produced precision of 82%, recall of 79%, and F1 score of 80.5%. This particular score highlights the effectiveness of the model in grasping the relevant terminologies excelled in the initial text.
- **ROUGE-2 (Bigram)**: The ROUGE-2 score was found to be lower than the previous metric, with a precision of 76%, a recall of 73%, and an F1 score of 74.5%. This implies that some two-word phrases or connotations created by a sentence could not be captured.
- **ROUGE-L (Longest Common Subsequence)**: It was seen that the F1 score was highly fulfilled by a score of 79% for the ROUGE-L indicating that the model was able to preserve the order of information that was represented within its original text.

These scores are presented in Table 2 below:

ROUGE Score Analysis:

ROUGE Score Type	Precision (%)	Recall (%)	F1 Score (%)
ROUGE-1	82.0	79.0	80.5
ROUGE-2	76.0	73.0	74.5
ROUGE-L	81.0	77.0	79.0

Figure 11: ROUGE Score Analysis

Model performance metric:

Fig 12: Model performance metric graph

The set of figures present the performance metrics of the model in terms of accuracy, precision, recall and F1 score depicted in the bar graph in Figure 12. These metrics help in understanding the performance of the model visually and each performance measure explain its proficiency in ensuring a proper trade-off in retrieval of correct information.

All above information we interpreted in table of Figure 11 in which all score type and its percentage.

10. Conclusion and Future Scope:

The Marathi text summarization project successfully demonstrates the application of natural language processing and machine learning techniques to generate concise and informative summaries from Marathi news articles. The implementation of the TextRank algorithm, combined with robust preprocessing steps, enabled the system to identify and extract key sentences, thereby producing meaningful summaries. The development of a user-friendly Streamlit application further enhanced the accessibility and usability of the tool, allowing users to effortlessly interact with the system and obtain summaries in real time. Overall, the project highlights the potential of automated text summarization to facilitate efficient information consumption and comprehension in regional languages like Marathi. Looking forward, there are several avenues for enhancing the capabilities and performance of the summarization tool. Integrating advanced machine learning models, such as transformers and BERT-based architectures, could significantly improve the accuracy and coherence of the generated summaries. Additionally, expanding the dataset to include a wider variety of topics and sources will help in training more generalized models, thereby increasing the tool’s versatility. Further, incorporating features like summarization customization, where users can specify the desired length or focus of the summary, can make the tool more adaptable to diverse user needs. Continuous feedback collection and iterative improvements based on user experiences will be crucial in refining the system, ensuring it meets the evolving demands of its users.

10. References:

1. Kale, S. D., Prasad, R., Potdar, G. P., Mahalle, P. N., Mane, D. T., & Upadhye, G. D. (2023). A Comprehensive Review of Sentiment Analysis on Indian Regional Languages: Techniques, Challenges, and Trends. International Journal on Recent and Innovation Trends in Computing and Communication, 11(9s), 93–110. https://doi.org/10.17762/ijritcc.v11i9s.7401

2. Kale, S. D., Mahalle, P. N., Kachhoria, R., Kumar, S., Chaudhari, P., & Patil, V. D. (2023c). Marathi text summarization through NLP and deep learning mechanism. Journal of Autonomous Intelligence, 6(3). https://doi.org/10.32629/jai.v6i3.1009

3. Dongare, M. (2022b). Present Status of Green Library Aspects and Its Implementation in Academic Libraries of Maharashtra. International Journal of Research in Library Science, 8(2), 99. https://doi.org/10.26761/ijrls.8.2.2022.1538

4. Abuobieda, A., & Osman, A. H. (2020b). An Adaptive Normalized Google Distance Similarity Measure for Extractive Text Summarization. Research gate, 163, 1–4. https://doi.org/10.1109/iccis49240.2020.9257668

5. Kale, S. D., & Prasad, R. S. (2018b). Author Identification on Imbalanced Class Dataset of Indian Literature in Marathi.

International Journal of Computer Sciences and Engineering, 6(11), 542–547. https://doi.org/10.26438/ijcse/v6i11.542547

6. Kale, S. D., & Prasad, R. S. (2020c). Influence of Language-Specific Features for Author Identification on Indian Literature in Marathi. In Advances in intelligent systems and computing (pp. 639–652). https://doi.org/10.1007/978-981-15-2475-2_59

7. Jain, A., Arora, A., Morato, J., Yadav, D., & Kumar, K. V. (2022b). Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm. Applied Sciences, 12(13), 6584. https://doi.org/10.3390/app12136584

8. Giri, V. V., Math, D., & Kulkarni, D. (2016b). A Survey of Automatic Text Summarization System for Different Regional Language in India. Bonfring International Journal of Software Engineering and Soft Computing, 6(Special Issue), 52–57. https://doi.org/10.9756/bijsesc.8242

9. Dheer, N., & Kumar, C. (2016b). Automatic Text Summarization: A Detailed Study. International Journal of Science and Research (IJSR), 5(2), 429–433. https://doi.org/10.21275/v5i2.nov161195

10. Deepali K. Gaikwad , C. Namrata Mahender: A Review Paper on Text Summarization.. International Journal of Advanced Research in Computer and Communication Engineering . 2016, 5:154. 10.17148/IJARCCE.2016.5340

11. Ab, A., & C, S. (2020b). An Overview on Document Summarization Techniques. International Journal of Recent Advances in Engineering & Technology, 08(03), 31–36. https://doi.org/10.46564/ijraet.2020.v08i03.007