Journal of Autonomous Intelligence (2023) Volume 6 Issue 3
doi: 10.32629/jai.v6i3.1009

ORIGINAL RESEARCH ARTICLE

Marathi text summarization through NLP and deep learning mechanism

Sunil D. Kale*, Parikshit N. Mahalle, Renu Kachhoria, Santosh Kumar, Prasad Chaudhari, Vivek D. Patil

Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Information Technology, Pune 411048, India

*** Corresponding author:** Sunil D. Kale, kalesunild@gmail.com

ABSTRACT

Every day, an ever-increasing amount of people gain access to the internet platform. This has proven to be efficient in creating cost-effective internet platform deployments and applications. The growth in the amount of people using the platform has resulted in a rise in the quantity of information accessible on the internet in the form of news, media, and other forms of communication. This causes evaluating and comprehending a significant amount of textual information a very challenging task. For the objective of generating textual summaries for Marathi texts, an effective and trustworthy approach is required. Through the use of machine learning methods, a successful strategy for extracting summary for the Marathi text has been generated for this objective. To obtain the Marathi text summary, the proposed method uses feature extraction as well as deep belief networks and decision tree methodologies. The experimentation was carried out on the performance of the Term Frequency-Inverse Document Frequency (TF-IDF) in the stopword elimination procedure, along with the evaluation of the summarization outcome which achieves a Mean Absolute Error (MAE) of 2.8 for the stopword removal approach through TF-IDF technique and a precision of 95.49% with an accuracy of 92.76%.

Keywords: natural language processing; TF-IDF; deep belief network; decision tree

ARTICLE INFO

Received: 19 July 2023
Accepted: 7 August 2023
Available online: 13 September 2023

1. Introduction

Knowledge and its pursuit have been one of the most basic driving forces behind the major improvements and other technological advancements across the globe. The preservation of past research and other documents has been significant in developing and growing our civilization on this planet. The current scientists and researchers build up their work on the past discoveries and researches. This is possible due to the various researches that have been available for studying, evaluation and enhancement through archiving and safeguarding the information. Nowadays, the internet platform allows for a large amount of information being available at the fingertips, which has facilitated the break neck speed of the transition and research in an extensive number of fields and paradigms. The easy availability of extensive knowledge about a certain field is one of the cornerstones for the development of the internet platform. The internet platform allows for the storage and distribution of large amounts of information to the public. The internet platform has been considerable in the realization of information sharing for a large number of researchers and reduced the time taken for information dissemination and research completion considerably. The internet has been evolving ever since, with more and more information being added to the platform every second. This has

been beneficial to a greater degree and has been the source of valuable information for many individuals.

The information explosion which is happening currently also has a number of repercussions, as the amount of information has been increasing exponentially every day, there is a chance that the information presented can be false or irrelevant. This is a real concern as there is a large amount of information that can be considered as erroneous which can lead to a lot of problematic scenarios. Any information can be better than no information that only holds true if the information is not compromised. This can create confusion that can be detrimental and extremely difficult to determine the authenticity and the accuracy of the information being present on the various web pages and other online resources.

There is also a lot of extensive and descriptive content on the internet, such as media and other educational resources as a result, there is a need for a convenient and practical text summarizing technique that can retrieve important information from a vast quantity of text while reducing the individual personal effort in obtaining relevant knowledge. Researchers have put a lot of effort into developing a text summarizing platform that can generate summaries for any text. However, discrepancies and poor precision have plagued these methods, making them unsuitable for this purpose.

In this research article, an innovative approach has been presented for achieving highly comprehensive and appropriate text summaries for Marathi texts by utilizing Natural Language Processing (NLP). The NLP framework is extremely beneficial since it enables the efficient retrieval of relevant summaries via a variety of procedures that are especially designed to help the computer comprehend the semantics of the text. Natural language processing has been combined with the use of linear regression to remarkable effectiveness with the inclusion of the deep belief network, the summary generating technique has been significantly upgraded. The deep belief network is in charge of implementing distinct activation functions including hidden layer estimation in order to enable machine learning deployment and enhance the consistency of the derived summaries. Through the use of the decision tree technique, the produced summaries are successfully categorized. The decision tree does this by employing if-then rules to classify the probability deriving from the deployment of a deep belief network in an accurate and comprehensive manner. The categorization yields a highly precise summary, which is described in depth in this research article.

Section 2 of this research paper works on past work under the name literature survey. Section 3 explains all the methodology in detail, whereas section 4 works on the evaluation of the results. Finally, section 5 concludes this paper and leaves behind some traces of future work.

2. Literature survey

Madhuri and Kumar elaborated automatic summarization of text is a multi-step process with many optional tasks. Every component has the capability of producing high-quality summaries. Identifying essential sentences from the provided material is a crucial element of the extractive summarization of text[1]. The authors suggested extractive text summarizing utilizing a statistical new technique focused on sentence ranking, with the summarizer selecting the phrases. The obtained phrases are turned into a summary text that is then transformed into audio. When compared to the traditional methodology, the suggested model improves accuracy. Agrawal presented that in the perspective of text mining, text summarization is a cutting-edge field[2]. Technological advances are being created all the time to make the procedure of summarizing easier. There are certainly additional and fresh opportunities for creating technology to reduce human effort in summarizing. Some novel techniques can be created to aid in the summarization of video clips and conversations into text form in the format expected. For individuals who are incapable of listening, new technology may be designed that will immediately transform the provided data, whether in the form of audios, videos, or text, summarize them, and instead transform them back into digital audio.

Ren and Guo elaborate that in NLP, text summarization is a quintessential activity[3]. The goal of text summarization automatically is to reduce the length of long texts. The neural networks centered on seq2seq with recognition have been proven to be effective in obtaining summaries. However, ensuring the correctness of the summary is challenging. Researchers introduced a text summarization approach using a copy mechanism and hybrid global gated unit in this paper. The method incorporates attentive architecture and the standard seq2seq algorithm. Experimentation on the datasets demonstrates that the technique improves assessment precision significantly when opposed to the previous benchmark algorithms. Abujar et al. demonstrated that when dealing with a text document, word embedding is crucial[4]. The lexicon of a written document is carried by all vectors. It groups all comparable words into a feature space and evaluates how similar the words are to one another. The authors in this article incorporate Bengali content that they found on the internet. Bengali text and its synopsis are included in the researchers' dataset. The authors were capable of improving the Bengali word embedding file that was used to deploy the dataset. The result of the provided strategy is superior. The authors intend to create a large dataset and wish to strengthen their Bengali language NLP capabilities.

Jadhav et al. introduce a text summarizer, which is designed to assist individuals such as students, office employees, and others[5]. Both abstractive and extractive approaches are used to condense an input text into an informative summary cosine similarity, stop words removal, TF-IDF, Longest Common Substring (LCS), and other extractive methodology methods were utilized. The authors programmed RNNs to alter and reduce the content and produce a relevant summary using abstractive approaches. The researchers used NLP and Natural Language Generator (NLG) to train the algorithm to comprehend text and convert it into organized information. NLP interprets and analyses human language by using machine learning and deep learning algorithms. Kale and Prasad demonstrated use of language explicit characteristics provides great accuracy results[6]. Abuobieda and Osman elaborated that the empirical approach is used to a text summarizing problem in the study[7]. To enhance the process of phrase grouping, the differential evolution algorithm was employed as a text segmentation approach. The chosen probabilistic model, Natural Gradient Descent (NGD), was effectively applied in a large dataset including hundreds of websites rather than a limited amount of sentences, as in a text with simple phrases. The typical single paragraph dataset was used as a test-bed dataset. In comparison to our new suggested technique, the application of the NGD measure of similarity resulted in poor subject diversification and coverage.

Masum et al. narrate that measurement of sentence similarity in Bengali can be done in a variety of ways[8]. The researchers used several algorithms and came up with a fantastic result, calculating the distance between two Bangla summary lines based on their similarities. This statistic aids in determining the degree of similarity between the machine-generated and human-generated summaries. Because abstractive summarization methods produce a word that is not always present in the supplied text content. It will be simple to designate a superior abstractive text summarization approach for the Bengali language if the comparison of the summary of supplied texts and machine output sentences can be established. Boorugu and Ramesh presented that some of the most impressive achievements in the field of text summarizing[9]. For many years, summarization has been a requirement because of the massive volume of information that is published on the internet every day. This article outlined all of the key summarizing approaches as well as the most recent research on each methodology. A singular text summary achieved high performance when contrasted to multi-document summarization, and a domain-specific summation attains better reliability in comparison to an approach that has no previous understanding of the subject.

Kale and Prasad demonstrated effect of imbalanced data on the accuracy result for author identification task on Marathi Language[10]. Talukder et al. elaborated that text summarization or content brief description is not an easy system to implement, however when compared to automated text summary systems, the extractive approach is a relatively simple one to implement[11]. In the extractive technique, the summary is created by

extracting the most important words or phrases from the provided material. Three works on abstractive summarization were evaluated by the authors. Three articles use various methods and mechanisms in their research. However, all of the articles showed a positive outcome. However, work on automated text summarization is still ongoing. Many studies are based on automatic text summarizing in order to build a model that can provide a precise summary. Janjanam and Reddy presented a basic examination of ideas, techniques, and algorithms related to automated summarization is presented in this work[12]. Some key components about text characteristics and their interpretations are provided as a prelude to the text summary. Though it is impossible to provide all of the work on summarizing methodologies, this research looked at both old and new material on techniques and strategies for automatic summarization. Primarily, this article explained machine learning approaches; however, it goes on to explore additional strategies for extractive and abstractive summarization utilizing optimization, semantic-based, and graph-based approaches. According to a study on deep learning textual summarizing techniques and technologies, several neural networks inspired methods are utilized for abstractive summarization[13]. The authors only deploy four methods in this study, and according to ROUGE F-scores, the pointer-generator using the overage technique model has the greatest ranking and produces the best performance when applied to the other systems. Amidwar et al. presented text analysis for author identification for Marathi[14]. Alfarra et al. presented Graph-based Fuzzy Logic and Extractive Text Summarization (GFLES) as a novel graph-based automatic summary generation method[15]. GFLES uses Document Graph Model (DGM) to retrieve phrase characteristics and Graph-based Growing Self-Organizing Map (G-GSOM) to group phrases in an attempt to determine text subtopics, after which Fuzzy Logic (FL) is used to score and rank phrases. GFLES also creates extractive summaries of single or several texts with a concentration on text subtopics. The framework was validated on a dataset, with the outcomes contrasted to two other comparable techniques one for individual documents and the other for numerous texts in terms of consistency and scope of the subtopics, the findings revealed that GFLES beat the other techniques. Hanunggul and Suyanto narrate that the global attention-based strategy outperforms the other models by generating additional phrases in the exact summary[16]. Kale et al. presented various features that can be considered for the sentiment analysis for Indian regional languages including Marathi[17]. However, because the technique of local attention analyses a subset of supplied input rather than the entire input words, the local attention-based method performs better, generating additional pairs of phrases in the actual summation. The dataset comprises a lot of characters and unfamiliar sentences that aren't mentioned in the word embedding dataset because it's presented in informal language. As a result, the performance rating does not exceed that of the standard English textual framework. For both algorithms, resetting all settings may result in greater results.

Digamberrao and Prasad presented that no consistent technique exists to determine the authorship and the Support Vector Machine (SVM) is superior throughout many situations[18]. For Author Identification (AI), there's a vast range of characteristics. It is evident that choosing of the features is a key responsibility in establishing an author's sentence structure. Selecting the feature relies on the subject and textual language. Functional set size is also significant and academics have attempted to optimize functions. This means that the authors will acquire fewer phrases from a certain author and it will eventually be a hard process in comparison with big textual information when recognizing the word choice of the specific author in order to determine literature in different linguistic writings. Digamberrao and Prasad established a new approach of authorship, which uses extraction of features and artificial intelligence methods in collaboration with statistic similarity models[19]. It's designed for several articles published by different contributors in the literary text in Marathi. For the development of the technical method, the Marathi conceptual papers evaluated by Sequential Minimal Optimization with Rule-based Decision Tree (SMORDT) (Rule-Based Decision Tree). Upon this basis of reminder measurement, precision, F-measure and exactness, the functionality of the suggested technique is assessed. Kale and Prasad presented review for the identification of authors based on their written literary

texts[20]. Recognizing the writer is to give the researchers a content of the same collection of publications as the authors, and facilitate the identification as word count, or, word wealth, special characters, the number of times the particular word, punctuation, connections, pronunciations, etc. The author's identification procedure requires providing training samples as text of different writers. This is the basis of the next step: categorize the writers and then provide the input textual data, i.e., test cases which should correspond to one of those writers who have been employed as a training set, but it's not the same textual information as already used in training.

3. Proposed methodology

The presented approach for the extraction of summaries of Marathi documents is discussed with the steps mentioned below. The **Figure 1** illustrates the system overview for the same.

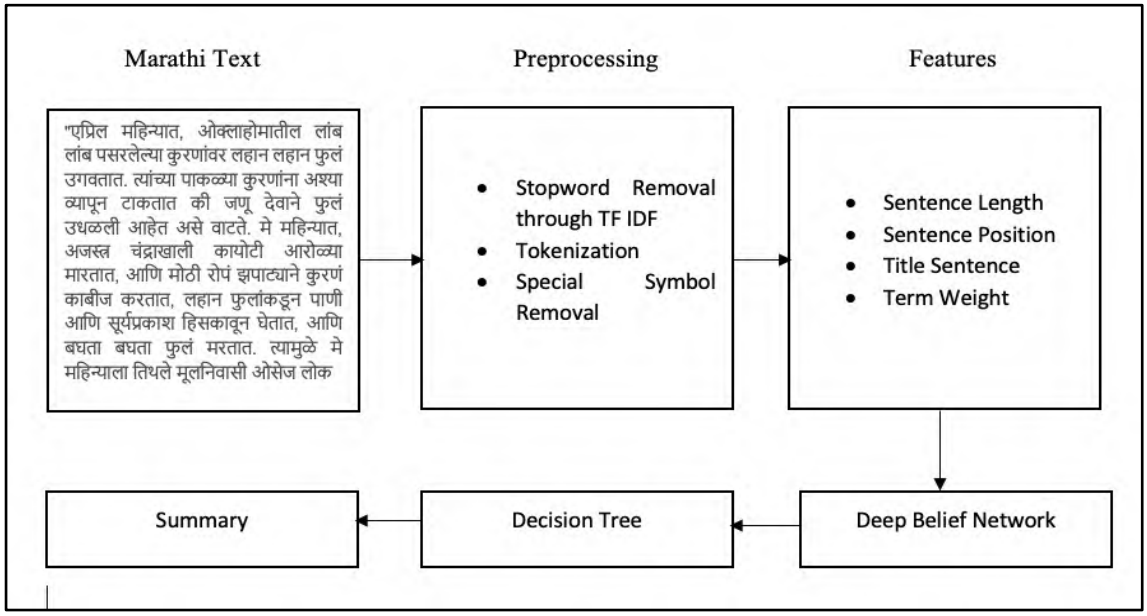

Figure 1. System overview of the Marathi document summary extraction.

Step 1: Document input—this is the initial step in the process, in which the system is given the Marathi document to summarize as an input. The document can be given in a variety of formats, including .doc, .pdf, and .txt. These files are saved in a specific folder and fed into the system as input. The input document data is retrieved in the form of string and this string is subjected to the further process of preprocessing as narrated in the next step.

Step 2: Preprocessing—the preprocessing stage utilizes the string extracted from the source text as an input. The preprocessing is done in order to ensure that the system is implemented correctly and effectively. There is unnecessary text in the source, which can considerably exacerbate the complexity of the method. As a result, data cleaning is required to eliminate these potentially harmful instances. The preprocessing approach has been described in detail in the following steps:

Special symbol removal—the special characters in the string are essential for punctuation while writing or conversing, but they aren't as beneficial in our method. Thus, the special symbols for example full stop (.), comma (,), semi-colon (;), colon (:) are removed from the input string in this step of preprocessing.

Tokenization—following the removal of the special characters, the string is sent to this stage of the preprocessing for tokenization. Tokenization is the process of converting a sequence of words into a well-indexed string. This transformation aids in the efficient use of this string in the system's following processes.

Stopword removal—stopwords in Marathi aren't as easy to identify and delete as they are in English. To

accomplish reliable and meaningful stopword detection, a natural language processing technique known as TF-IDF is used. TF-IDF is an acronym for term frequency and inverse document frequency. This method determines the relevance of certain terms as well as the quantity with which they are used. This method is used to delete the least significant terms from the list, which are often stopwords.

The elimination of these stopwords is beneficial since they offer no additional meaning to the phrase and greatly increase execution speed. In this technique all words are estimated for their respective TF-IDF values. Then they are sorted in ascending order to get the top ten words to consider them as stop words. After this process these words are then eliminated on their occurrence in the text. The TF-IDF can be measured using the Equation (1) mentioned below.

$$TF - IDF = TF \times \log \text{ (number of documents/number of documents containing word W)} \quad (1)$$

Step 3: Feature extraction—the feature extraction procedure is the most fundamental element of the system since it successfully identifies and extracts the necessary elements. This step utilizes the preprocessed string for performing the processing. The complete procedure is illustrated below, along with the different features that are retrieved.

The entire procedure can be illustrated in detail in the Algorithm 1 given below.

Algorithm 1 TF-IDF estimation

1: Step 0: Start
2: Step 1: Read the input string
3: Step 2: Divide string into words on space and store in a vector SV
4: Step 3: for i = 0 to N (where N is the length of SV)
5: Step 4: WC = SV [i]
6: Step 5: Count WC for the respective input text as TF
7: Step 6: Count WC for the all other texts that as DF
8: Step 7: IDF = log (DF)
9: Step 8: TF-IDF = TF × IDF
10: Step 9: end for
11: Step 10: Stop

The title sentence is the actual opening line in the text, and it serves as the introduction. As a result, the remainder sentences are contrasted to the title sentences in order to obtain this feature, as shown in Equation (2).

$$TF = Frequency\ of\ title\ sentence\ words\ in\ the\ sentence/Sentence\ length \quad (2)$$

Sentence length—the length of the sentences is another essential characteristic that must be assessed. In contrast to the other sentences, the sentence length gives much more detail. The following Equation (3) is used to measure this feature.

$$SLF = Sentence\ length/Biggest\ sentence\ length \quad (3)$$

Term weight—this is really a feature that is unique to a certain sentence. The quantity of significant words in a phrase is referred to as weight. The greater the word weight, the more significant the statement is deemed. Equation (4) below is an excellent way of extracting this feature.

$$TWF = Frequency\ of\ top\ 10\ words\ in\ a\ sentence\ of\ the\ document/Sentence\ length \quad (4)$$

Sentence position—the sentence's position is also a useful indicator for determining the sentence's significance. The value of a sentence is determined by its location. This metric is attained by assigning a score to each sentence in the prescribed sequence. The first phrase gets a 1, the second gets a 0.8, the third gets a 0.6, the fourth gets a 0.4, and the fifth gets a 0.2. After the fifth, the remaining sentences are given a score of 0.

Step 4: Deep belief network—the feature list from the previous phase is used as an input for the development of deep belief networks. This feature list is successfully arranged in decreasing order, allowing the extraction of the target values like 1 and 0, which are the higher and lower values, respectively. The number

of features retrieved in this methodology is 4, thus 16 random weights are applied to these characteristics.

The resulting data may be used to calculate the probability score of the deep belief network. The resulting values are segregated according to the activation function "tan h" to produce the probability score. These values are ultimately recorded in the form of a list and added to the DBN's model. The following phase successfully provides the trained model comprising the values of the probability scores for the intention of appropriate categorization via decision tree deployment.

$$\tan h = (e^{x} - e^{-x})/(e + e) \tag{5}$$

where, x are the values of the input attributes.

The process for hidden layer evaluation has been shown in the Algorithm 2 given below.

Algorithm 2 Hidden layer evaluation

```
1: //Input: Feature List FL, Weight set WSET = { }
2: //Output: Hidden Layer value list HVL hidden Layer Evaluation (FL, WSET), index = 0
3: Start
4: HLV = ∅ {Hidden Layer value]
5: for i = 0 to size of FL
6: ROW = FL [i]
7: for j = 0 to size of ROW
8: X = 0
9: for k = 0 to N [Number of Neurons]
10: ATR = ROW [j]
11: X = X + (ATR × WSET [index])
12: index ++
13: end for
14: HVL = tan h (X)
15: end for
16: end for
17: return HVL
18: Stop
```

Step 5: Decision tree—this is the most essential component since it is where the system generates the final summary. This is accomplished by segregating the probability scores generated by the deep belief network's output in ascending order. The decision tree technique is implemented using if-then rules that categorize the probability scores entirely and only enable the most comprehensive summaries to be produced. The user selects the level of the summary by providing a percentage of input text to be summarized. The final outcome is a textual summary that can be seen in the IDE user interface or saved in an output file.

4. Results and discussions

This technique was subjected to an experimental investigation in order to assess the effectiveness of the system. The Mean Absolute Error (MAE) and precision and recall are the performance statistics that are used for the evaluation purpose. The Mean Absolute Error accurately detects and evaluates the errors that are produced by the presented framework. The precision and recall metrics allow for the correct understanding of the execution accuracy of the approach. The next section adequately elaborates on the experimental findings.

5. Performance evaluation based on Mean Absolute Error

The Mean Absolute Error has been incorporated in order to realize the tendency for error that is generated by the suggested system. The Mean Absolute Error is used to quantify the system's error tendencies in percentage format. For the objective of determining the system's MAE, the Equation (6) is being used below. The use of continuous characteristics is required for this assessment, and the term frequency and inverse document frequency TF-IDF are the properties that are considered. This NLP technique is used to locate stopwords in the source Marathi text and remove these for preprocessing purposes.

$$MAE = \sum_{i=1}^{n} |Yi - Xi|/n \quad (6)$$

where, *Xi*: number of to be eliminated stopwords; *Yi*: number of actual stopwords through TF-IDF; *n*: number of experiments conducted.

A total of ten trials are carried out with the goal of attaining the *MAE* for the detection and eradication of stopwords. In **Table 1** below, the experimental results are presented in a tabular format. **Figure 2** shows a visual representation of the experimental results.

Table 1. Experimentation and calculation of MAE.

Experiment number (*n*)	Number of eliminated stopwords (*Xi*)	Number of actual stopwords (*Yi*)	Difference (*Xi*−*Yi*)
1	13	10	3
2	19	15	4
3	12	11	1
4	10	8	2
5	19	18	1
6	16	14	2
7	23	19	4
8	21	18	3
9	16	13	3
10	22	17	5
-	-	*MAE*	2.8

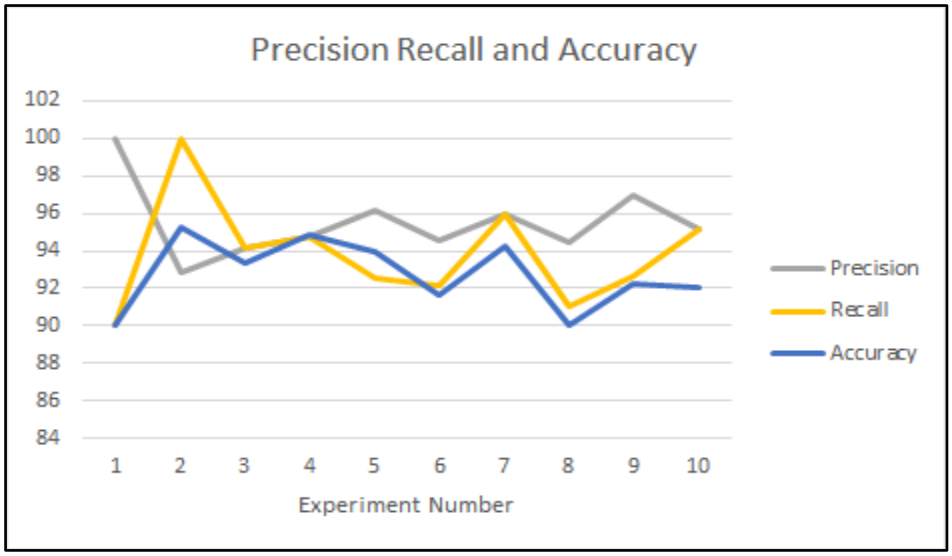

Figure 2. Graphical representation of the precision, recall and accuracy values.

6. Performance evaluation through precision and recall

Amongst the foremost and significant assessment attributes that extracts the system's effectiveness is precision and recall performance measures. The precision is governed by the model's relative accuracy. Recall, on the other hand, may be thought of as the model's absolute consistency. The Accuracy is calculated by taking the aggregate of all the examples evaluated throughout the examination. This Accuracy score can also be computed using Equation (9) shown below, along with precision and recall through the Equations (7) and (8) respectively.

$$Precision = A/(A + B) \quad (7)$$

$$Recall = A/(A + C) \quad (8)$$

$$Accuracy = (A + D)/(A + B + C + D) \quad (9)$$

where, A = No. of correct sentences selected for the summary; B = No. of incorrect sentences selected for the summary; C = No. of correct sentences not selected for the summary; D = No. of incorrect sentences not selected for the summary.

The precision and recall metrics were used to evaluate the Marathi text summarization module's performance. The results of the comprehensive testing done on the mechanism for a number of trials with escalating numbers of sentences delivered with each iteration are summarized in **Table 2** below. The Marathi text summary outcomes were also displayed in the graphical form shown **Figure 2**.

Table 2. Precision, recall and accuracy performance.

Experiment number	No. of sentences	No. of correct sentences selected	No. of incorrect sentences selected	No. of correct sentences not selected	No. of incorrect sentences not selected	Precision	Recall	Accuracy
1	10	9	0	1	0	90	100	90
2	20	13	1	0	6	100	92.85	95
3	30	16	1	1	12	94.11	94.11	93.33
4	40	19	1	1	19	95	95	95
5	50	25	1	2	22	92.59	96.15	94
6	60	35	2	3	20	92.1	94.59	91.66
7	70	47	2	2	19	95.91	95.91	94.28
8	80	51	3	5	21	91.07	94.44	90
9	90	63	2	5	20	92.64	96.93	92.22
10	100	79	4	4	13	95.18	95.18	92

The precision, recall, and accuracy scores tested for the Marathi text summarizing ability demonstrated the technique's deployment quality of the summaries. Precision is 93.86%, recall is 95.51% and accuracy is 92.75% using this approach. These results are highly encouraging for an initial execution of a system for Marathi text summary creation that achieves the precision, recall, and accuracy levels listed above.

These figures show how the performance standards of the Marathi text summarizing technique proposed in this study were effectively accomplished. This technique is contrasted with the technique, which uses a latent semantic analysis strategy to summarize Kannada text[21]. The results are reviewed and summarized in **Table 3** below, as well as a graph in **Figure 3**.

Table 3. Precision, recall and accuracy performance comparison.

Methodology	Precision	Accuracy
LSA based approach	65	87
Proposed approach	93.86	92.75

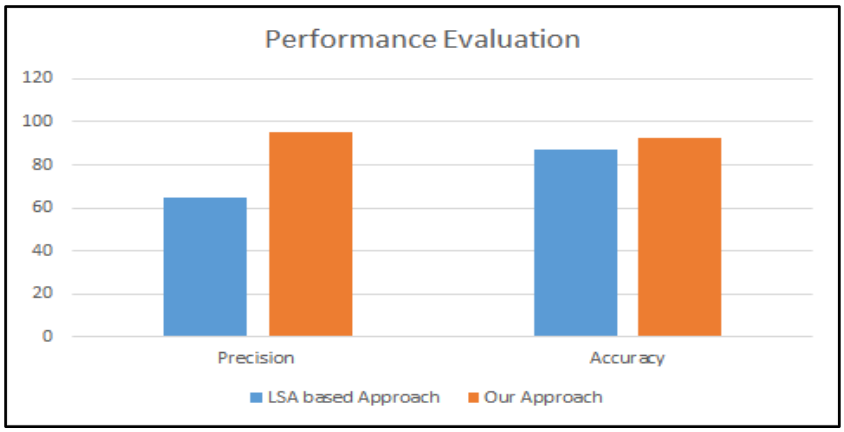

Figure 3. Graphical representation of the precision, recall and accuracy values.

Our technique outperforms by a wide margin, as seen in the performance comparison above. This is a very positive result that illustrates the attainment of the improvements made possible by our technique.

7. Conclusion and future scope

In this research paper, the proposed approach for the creation of text summary of Marathi texts is discussed. There have been very few studies dedicated to producing a text summary for Marathi text files. The Marathi text is used as an input in this summarization method, and it is adequately preprocessed to condition the data. Stopword elimination, tokenization, and special symbol removal are the three processes used to accomplish this. Stopwords are removed using the term frequency and inverse document frequency techniques, with the least significant words being identified and deleted. The feature extraction procedure is applied to the preprocessed string. Sentence length, sentence position, title sentence, and term weight are the characteristics retrieved in this stage. The retrieved features are gathered and fed to a deep belief network for evaluation. The characteristic features of the text are input into the deep belief network, which generates a useful summary in the format of probability score. To obtain an effective and accurate summary, these scores are successfully separated in the form of a decision tree on the basis of the summary percentage selected by the user. The experimentation on the system has been effective in achieving a Mean Absolute Error of 2.8 for the stopword removal approach through TF-IDF technique. And the model achieves a precision of about 95.49% and accuracy of 92.76%.

For the future enhancement an improved web crawler can be built to acquire the Marathi text summary from the web sites of news agencies and other websites in the effort to enhance the suggested system.

Author contributions

Conceptualization, SDK and PNM; methodology, RK and SK; validation, PC and VDP. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Madhuri JN, Kumar RG. Extractive text summarization using sentence ranking. In: Proceedings of the 2019 International Conference on Data Science and Communication (IconDSC); 1–2 March 2019; Bangalore, India. pp. 1–3.
2. Agrawal K. Legal case summarization: An application for text summarization. In: Proceedings of the 2020 International Conference on Computer Communication and Informatics (ICCCI); 22–24 January 2020; Coimbatore, India. pp. 1–6.

3. Ren S, Guo K. Text summarization model of combining global gated unit and copy mechanism. In: Proceedings of the 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS); 18–20 October 2019; Beijing, China. pp. 390–393.
4. Abujar S, Masum AKM, Mohibullah M, et al. An approach for Bengali text summarization using Word2Vector. In: Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 6–8 July 2019; Kanpur, India. pp. 1–5.
5. Jadhav A, Jain R, Fernandes S, Shaikh S. Text summarization using neural networks. In: Proceedings of the 2019 International Conference on Advances in Computing, Communication and Control (ICAC3); 20–21 December 2019; Mumbai, India. pp. 1–6.
6. Kale SD, Prasad RS. Influence of language-specific features for author identification on Indian literature in Marathi. In: *Soft Computing and Signal Processings*. Springer, Singapore; 2020. pp. 639–652.
7. Abuobieda A, Osman AH. An adaptive normalized google distance similarity measure for extractive text summarization. In: Proceedings of the 2020 2nd International Conference on Computer and Information Sciences (ICCIS); 13–15 October 2020; Sakaka, Saudi Arabia. pp. 1–4.
8. Masum AKM, Abujar S, Tusher RTH, et al. Sentence similarity measurement for Bengali abstractive text summarization. In: Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 6–8 July 2019; Kanpur, India. pp. 1–5.
9. Boorugu R, Ramesh G. A survey on NLP based text summarization for summarizing product reviews. In: Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA); 15–17 July 2020; Coimbatore, India. pp. 352–356.
10. Kale S, Prasad R. Author identification on imbalanced class dataset of Indian literature in Marathi. *International Journal of Computer Sciences and Engineerin*g 2018; 6(11): 542–547.
11. Talukder MAI, Abujar S, Masum AKM, et al. Comparative study on abstractive text summarization. In: Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies. (ICCCNT); 1–3 July 2020; Kharagpur, India. pp. 1–4.
12. Janjanam P, Reddy CP. Text summarization: An essential study. In: Proceedings of the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS); 21–23 February 2019; Chennai, India. pp. 1–6.
13. Tandel J, Mistree K, Shah P. A review on neural network based abstractive text summarization models. In: Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT); 29–31 March 2019; Bombay, India. pp. 1–4.
14. Amidwar S, Baxi S, Rao K, Kale S. Text analysis for author identification using machine learning. *Journal of Emerging Technologies and Innovative Research* 2017; 4(6): 138–141.
15. Alfarra MR, Alfarra AM, Alattar JM. Graph-based fuzzy logic for extractive text summarization (GFLES). In: Proceedings of the 2019 International Conference on Promising Electronic Technologies (ICPET); 23–24 October 2019; Gaza, Palestine. pp. 96–101.
16. Hanunggul PM, Suyanto S. The impact of local attention in LSTM for abstractive text summarization. In: Proceedings of the 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI); 5–6 December 2019; Yogyakarta, Indonesia. pp. 54–57.
17. Kale SD, Prasad R, Potdar GP, et al. 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* 2023; 11(9s): 93–110. doi: 10.17762/ijritcc.v11i9s.7401
18. Digamberrao KS, Prasad RS. Author identification on literature in different languages: A systematic survey. In: Proceedings of the 2018 International Conference on Advances in Communication and Computing Technology (ICACCT); 8–9 February 2018; Sangamner, India. pp. 174–181.
19. Digamberrao KS, Prasad RS. Author identification using sequential minimal optimization with rule based decision tree on Indian literature in Marathi. *Procedia Computer Science* 2018; 132: 1086–1101. doi: 10.1016/j.procs.2018.05.024
20. Kale SD, Prasad RS. A systematic review on author identification methods. *International Journal of Rough Sets and Data Analysis (IJRSDA)* 2017; 4(2): 81–91. DOI: 10.4018/IJRSDA.2017040106
21. Geetha JK, Deepamala N. Kannada text summarization using latent semantic analysis. In: Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 10–13 August 2015; Kochi, India. pp. 1508–1512.