Text Summarization and Classification for Indian Language

Manasi Chouk
Department of Computer Engineering (M.E)
Shree L. R. Tiwari College of Engineering, Mumbai, Maharashtra, India

Neelam Phadnis
Department of Computer Engineering (M.E)
Shree L.R. Tiwari College of Engineering, Mumbai, Maharashtra, India

ABSTRACT
Over the last few years, there have been significant advances in Text Summarization. Text Summarization can be implemented using two approaches; one is the NLP based approach and another is Deep Learning approach. Text Summarization is a demanding and fascinating field of NLP. It has become important because of the tremendous increase in information and data. Text Summarization is technique of creating a specific and relevant short abstract of text using different ways like books, news articles, research papers, tweets etc. Research is being done to summarize large text documents which are difficult to summarize manually. For English and other foreign languages various automated text summarization systems are available. However very few techniques are available for Indian language such as Marathi. In this paper, two extractive techniques are proposed to summarize large Marathi texts. This paper also performs classification on Marathi text using Marathi headlines dataset.

General Terms
Text Summarization, Text Classification

Keywords
NLP, Extractive technique, TF-IDF, Text Rank, Marathi Language

1. INTRODUCTION
Natural Language Processing (NLP)is a branch of Artificial Intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines including computer science and computational linguistics..The motivation for NLP was the significant growth of textual data(huge amounts of unstructured data, webpages and social media, emails and documents).The goal of this field is to perform useful tasks involving human language, tasks like enabling human-machine communication(conversational agents or dialogue systems includingautomatic speech recognition),improving human-human communication(natural languageunderstanding), or simply doing usefulprocessing of text or speech(speech synthesis).Another useful task is that ofmakingavailable to non-English-speaking readers the vast amount of scientific information on the Web in English.

A summary is a short abstract of text which gives important information from the original document. The advantage of using a summary is that it reduces the readingtime. The aim of automated text summarization systems is to give a shorter version of input text document with semantics. The intention is to create a clear and easy to understand summary with main points from the document. There are two approaches used to summarize the text, Extractive and Abstractive on the basis of whether the exact sentences are considered as they appear in the original text.

An extractive summarization technique comprises of selecting relevant words, sentences etc. from the original document and combining them into a shorter form. Extractive summarizers are mostly based on scoring sentences from the input document. Recently, the most common techniques used either statistical or linguistic approach.

In Abstractive summarization, important concepts of a document are identified and then expressed those concepts in natural language. The abstractive algorithms creates new phrases and sentences that are relevant to theconcepts identified from the original document. Compared to extractive technique very less work has been done in this area for Indian languages.Abstractive technique is broadly classified into two types: Structured based and Semantic based.

In India, Marathi is one of the prominent regional languages.So far very less work has been done on Marathi language text summarization. This paper,uses two extractive techniques i.e. TFIDF and Text Rank algorithms to summarize the Marathi document. In extractive technique, the process has two steps first is Pre- processing step and second is Processing step. Pre-processing is a structured representation of original document which includes stop-word elimination, stemming etc. whereas in processing step features influencing the relevance of sentences are decided and calculated according to the algorithm.

2. RELATED WORK
In today's digital world, summarization methods are greatly needed to consume the ever-growing amount of text data available online It will become easier if the original text is converted into shorter form with semantics for the readers. This section describes the research work done in the field of text summarization and classification for Marathi language.

Sheetal Shimpikar and Sharvari Govilkar has reviewed and compared various text summarization techniques which can be used for indian regional languages and discussed in detail two types of text summarization techniques i.e extractive and abstractive. [1]

Virat V. Giri and et.al have described single document multi news Marathi extractive summarizer system. This extractive summarizer system is used to summarize the single Marathi document with multi news by preserving important sentences based on statistical and linguistic text features. [2]

Apurva D. Dhawale and et. al have explored pre-processing techniques for Marathi e-news articles. They have used

various algorithms for processing Marathi text such as Text ranking, LINGO, Supervised Learning Method, Clustering, lexical chain, and domain specific summarization algorithms [3]

Mudassar M. Majgaonker and et.al have presented and evaluated a rule based and an unsupervised Marathi stemmer. They have done suffix stripping which is pre-processing step and can be done using stemmer for Marathi text. [4]

Deepali K. Gaikwad and et. al presented a question based text summarization system using rule based stemmer. This rule based stemmer was used for generating an appropriate question from the given input. [5]

Shubham Bhosale and et. al proposed a text summarization system by extracting keywords from the e-news articles using two modules word extraction module and summarization module. [6]

Jayshri Arjun Patil and et.al have reviewed Name Entity Recognition system for Marathi language. They have discussed various issues, challenges, and approaches which can be used for NER Marathi system. [7]

Vaishali V. Sarwadnya and Sheetal S. Sonawane proposed a graph based technique for Marathi document. They have used two technique in Text Rank algorithm i.e positional and similarity. [8]

Nutan B. Zungre and et.al proposed a work using graph based algorithm by which word ambiguity has been resolved based on their meaning and context. [9]

Anishka Chaudhari and et. al proposed a extractive text summarizer using Neural network. They have used Recurrent Neural Network (RNN) a type of Neural Network which worked on sequential data. Also they have used Google translate API for translating Marathi text to English text. [10]

Pooja Bolaj and SharvariGovilkar presented a text classification systems using supervised learning and ontology based methods. They have used supervised learning methods which included Naïve Bayes (NB), Modified K-Nearest Neighbour (MKNN) and Support Vector Machine (SVM). [11]

3. TEXT SUMMARIZATION AND CLASSIFICATION SYSTEM FOR INDIAN LANGUAGE

3.1 Proposed Approach

In this paper automated text summarization system is proposed taking input as a Marathi text documentusing two techniques TFIDF and Text Rank method. In text classification system for an input Marathi dataset N-Gram creation is done followed by modelling techniques Logistic Regression and Neural network.

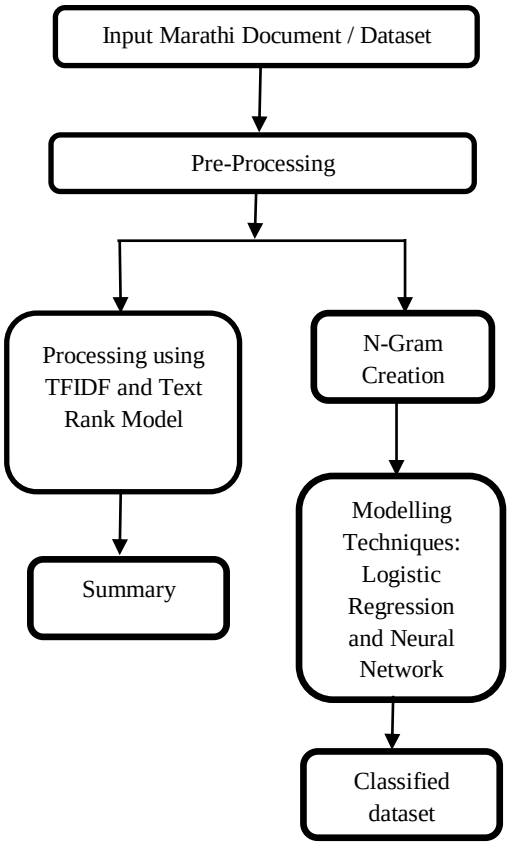

Fig.1: Text Summarization and Classification System for Indian Language

3.2 Text Summarization System

In the proposed system, for text summarizationextractive techniques are used to summarize large Marathi text and analyse the result obtained from them.

In this work, Marathi text document has been used as input to system. Marathi is a morphologically rich language. For implementation Python programming language has been used because in Python various libraries are available which can be used for developing NLP applications.

In Pre-Processing, three steps has been performed; first is the tokenization of sentence and then into words, second is the removal of stop words and the third step is stemming i.e. reducing the word to their stem or root.

After Pre-Processing the next step is the Processing step in which the TFIDF and Text Rank algorithmsare implemented.

3.2.1 Term Frequency - Inverse Document Frequency (TFIDF)

TFIDF is a numerical statistical method which is used to find out the importanceof a word in a set of documentsor corpus. In this model, meaningful information is collected and stored. From this information the uncommon words are identified and stored and these words are more important than common words. It is often used asa weighting factor in searches of information retrieval, text mining, and user modelling. TF of a word is defined as the term frequency of that word in that particular document and IDF of word in whole corpus of document.

TF=Number of Occurrences of a word in the document / Number of word in the document

IDF=Number of document / Number of document containing word.

The tf–idf value is directly proportional tothe number of times a word appears in the document. Resulting tf-idf value is divided by number of documents in the corpus that contain this word. By doing this adjustment is done for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today. The Pre-Processing steps are same for both the techniques. In TFIDF sk-learn feature is used. It is a software machine learning libraryfor the python programming language available on internet. It features classification, regression and clustering algorithm. This feature converts a collection of raw documents to a matrix of tf-idf features.

3.2.2 Text Rank

Text Rank is a text summarization technique which is used in Natural Language Processing to generate Document Summaries.Text Rank uses an extractive approach and is an unsupervised graph-based text summarization technique. Text Rank Algorithm is inspired by Page Rank algorithm which is primarily used for ranking web pages in online search results. In Text Rank algorithm similarity measure is used. A graph is created out of the sentences; while nodes represents sentences and the count on edges between two nodes is found out using similarity measure function. Using these function similar words between the two sentences is found out with their count iteratively until consistent word counts are obtained. Then the Page Rank algorithm is applied to the text with only difference is that nodes are the sentences instead of web pages. Finally the sentences are sorted in descending order and the top ranked sentences are chosen to be a part of summary.

3.3 Text Classification System

Text Classification is an application of Natural Language Processing by which we can create a model and that model is used to classify human language. In this model, N- Gram based Text Classifier for Marathi text has been implemented and used supervised learning methods to classify Marathi text using Logistic Regression and Neural Network Classifier. in this system, open source dataset from iNLTK i.e. indicnlp corpus has been used. [Marathi news headlines dataset]. iNLTK is Natural Language Toolkit for Indic Languages which consist of various predefinedlanguage modelswhichcan be used by NLP applications.

In the proposed model, Pre-Processing phase includes tokenization, stop word removal and N-Gram creation. Then the models /algorithm logistic regression and neural network classifier used to classify the data. Below is the description of the technic and models used.

3.3.1 N-Gram Modelling Technique

N-Gram Model is defined as "A contiguous sequence of N items from a given sample of text". The sequence of items can be character, word, or sentence, whereas N is a integer number.If N is 1 then it is Unigram, if N is 2, then that sequence is a bigram,if N is 3 then that sequence is a trigram, and so on. In this modelling technic, the probabilistic method is used to trained on the corpus of text. It predicts the most probable character or word from the given text of sequences. These models are used in various applications like predictive text input, machine translation and speech recognition. This model checks that how many times a word is repeated in given text and then calculates the probability of that word. This model is simple model and has some disadvantages which can be overcome by interpolation smoothing and back off. This language model finds the probability distribution over word sequence.

3.3.2 Logistic Regression

LR is a supervised learning classification algorithm which is used to predict the probability of a target variable.Mathematically, this model predicts P(Y=1)as a function of X. This is a simple ML algorithm and it has been used for different classification models such as spam detection, diabetes prediction etc.

3.3.3 Neural Network Classifier

Class MLPClassifier implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation. Multi-layer Perceptron (MLP**)** is a supervised learning algorithm. In this model, there can be one or more than one nonlinear layers between input and output layers. This layers is called as hidden layer. This function $f(\cdot):R^m \rightarrow R^o$ learns by training on a dataset, where m is the number dimensions for input and o is the output. For a set of features X=x1,x2,...,xm and a target y, Approximately, this function learns a nonlinear function for classification or regression.

4. EVALUATION AND RESULTS

In the proposed model of text summarization, the summaries are evaluated using ROUGE which stands for Recall-Oriented Understudy for Gisting Evaluation. It is a set of metrics for used to check the summaries generated. It compares an automatically generated summaries with areference summaries.

Recall in context of ROUGE means how much of the reference summary is the system summaries captures or recovers. This can be evaluated as –

Number of overlapping words /Total words in reference summary.

Whereas Precision means how much of the system summary is relevant or needed. It can be computed as-

Number of overlapping words/ Total words in system summary.

4.1 Results of Text Summarization System

In this work, summaries are evaluated using Rouge 2 which means overlapping of bigrams between the system and reference summaries.

Below are examples of the input file and summaries generated by text rank and tfidf algorithms.

1. Example of input file

पाककडूनपुन्हाएकदाशस्त्रसंधीचेउल्लंघन
जम्मूतीलपुँछजिल्ह्यातपाकिस्तानीसैन्यानेप्रत्यक्षनियंत्रणरेषेवरील (एलओसी)
भारतीयचौक्यांवरगोळीबारकरतपुन्हाएकदाशस्त्रसंधीचेउल्लंघनकेले श्रीनगर-
जम्मूतीलपुँछजिल्ह्यातपाकिस्तानीसैन्यानेप्रत्यक्षनियंत्रणरेषेवरील

(एलओसी)
भारतीयचौक्यांवरगोळीबारकरतपुन्हाएकदाशस्त्रसंधीचेउल्लंघनकेले.
शनिवारीरात्रीत्यांनीअंदाधुंदगोळीबारकेला.
पुँछजिल्ह्यातीलशहापूरसेक्टरमधीलभारतीयचौक्यांवरपाकिस्तानीसै
न्यानेगोळीबारकरण्याससुरूवातकेल्यानेभारतीयजवानांनीहीयागोळी
बारालाचोखप्रत्युत्तरदिले.
पाकिस्तानीसैन्याकडूनपहाटेसाडेचारपर्यंतगोळीबारसुरुहोता.
यागोळीबारातएकहीभारतीयजवानजखमीझालेलानाही,
अशीमाहितीलष्करीअधिका-यांनीदिलीआहे.
शनिवारीझालेल्यायागोळीबारामुळेअनेकपरिसरातशोधमोहिमसुरूक
रण्यातआलीआहे.

2. Example of summary generated by text rank algorithm

पाककडूनपुन्हाएकदाशस्त्रसंधीचेउल्लंघन
जम्मूतीलपुँछजिल्ह्यातपाकिस्तानीसैन्यानेप्रत्यक्षनियंत्रणरेषेवरील
(एलओसी)
भारतीयचौक्यांवरगोळीबारकरतपुन्हाएकदाशस्त्रसंधीचेउल्लंघनकेले
श्रीनगर-
जम्मूतीलपुँछजिल्ह्यातपाकिस्तानीसैन्यानेप्रत्यक्षनियंत्रणरेषेवरील
(एलओसी)
भारतीयचौक्यांवरगोळीबारकरतपुन्हाएकदाशस्त्रसंधीचेउल्लंघनकेले.
पुँछजिल्ह्यातीलशहापूरसेक्टरमधीलभारतीयचौक्यांवरपाकिस्तानीसै
न्यानेगोळीबारकरण्याससुरूवातकेल्यानेभारतीयजवानांनीहीयागोळी
बारालाचोखप्रत्युत्तरदिले.

3. Example of summary generated by tfidf algorithm

शनिवारीरात्रीत्यांनीअंदाधुंदगोळीबारकेला.
पाकिस्तानीसैन्याकडूनपहाटेसाडेचारपर्यंतगोळीबारसुरुहोता.

Below is the table showing rouge values of the summaries.

Table 1 Rouge values

Rouge Values - Precision		
Input	Tfidf	Text Rank
Document 1	0.9220	0.6227
Document 2	0.9405	0.5540
Document 3	0.8379	0.7147
Document 4	0.9066	0.8412
Document 5	0.7847	0.5278

Considering output summaries and precision score TFIDF technique is more relevant than Text Rank technique.

4.2 Results of Text Classification System

In Text Classification, N- Gram based text classifier is implemented to classify Marathi news headlines dataset using logistic regression and neural network classifier. Chosen dataset had three classes lifestyle, entertainment andsports. The total number of data elements present in data set is 3815. Out of 3815 data elements 1306 wereclassified as belonging to lifestyle, 1273 were classified asbelonging toentertainment and 1236 were classified as belonging to sports.

The Accuracy of neural network classifier is 96.02 % and logistic regression is 94.76% which is satisfactory. In this case, n-gram value is considered as (1,1) i.e. unigram. As n gram model predicts the most probable word that might follow the sequence . Also below are the results of the models that have been shown using confusion matrix.

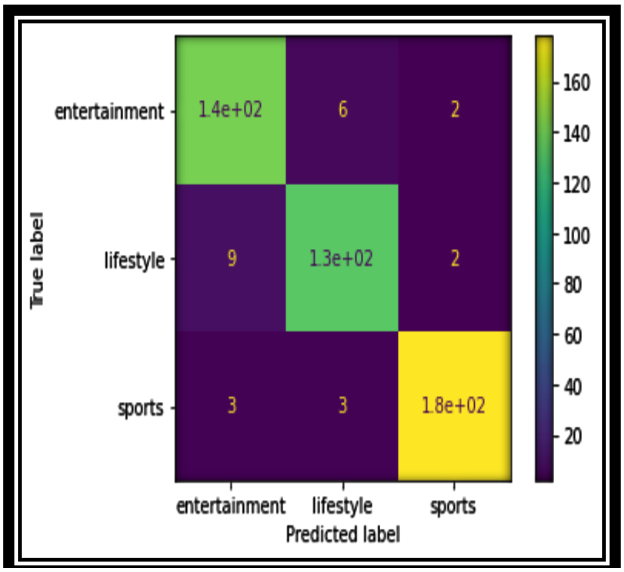

Fig.2: Confusion matrix using logistic regression

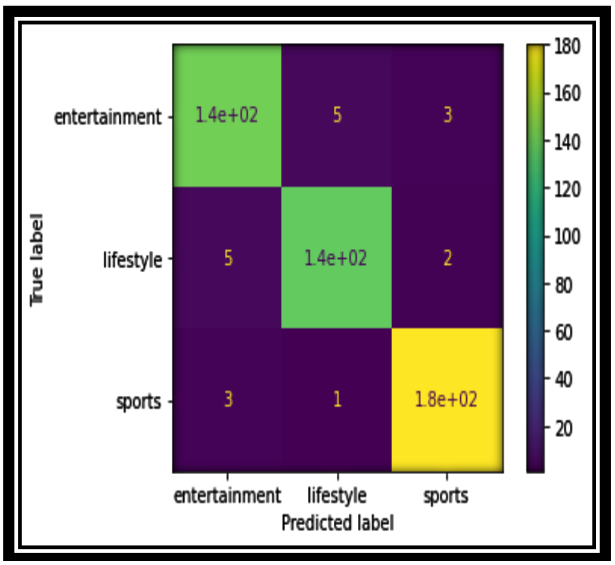

Fig.2: Confusion matrix using MLP classifier

Another aspect in this system is to check the accuracies of the model by changing the n gram values for Marathi text which means by changingvalues from unigram to bigram, trigram and so on. Below is the table which shows the accuracy of both the models after changing the n gram values.

Table 2 Accuracy of MLP and logistic regression

	N – gram Range	Model / Algorithm	Test Accuracy
1	(1, 1)	Logistic Regression MLP Classifier	94.76% 96.02%
2	(1, 2)	Logistic Regression MLP Classifier	94.14% 95.60%
3	(1, 3)	Logistic Regression MLP Classifier	93.93% 96.65%
4	(1, 4)	Logistic Regression MLP Classifier	93.93% 95.81%

5	(1, 5)	Logistic Regression MLP Classifier	93.93% 95.81%
6	(2, 2)	Logistic Regression MLP Classifier	66.73% 65.48%
7	(3, 3)	Logistic Regression MLP Classifier	31.59% 31.59%

The above table shows the different n gram values with lower and higher boundary range of the word to be extracted. The default range is (1, 1) means unigram, (1, 2) means unigram and bigram, (1,3) means unigrams, bigrams and trigrams, (2, 2) means only bigrams and so on.The result shows that by considering two words(bigram) or three words (trigram) in a sequence of words the probability of predicting the next word by the algorithm is poor.

5. CONCLUSION

In this paper, two systems were proposed for Marathi text namely, text summarization and text classification. It was a challenging task, because very less work has been done in the field of text summarization and classification for Marathi text. Marathi is a morphologically rich language and most of the lexical methods required for pre-processing had to be implemented from scratch.

In text summarization system, extractive technique is used and compared the summaries. We can conclude that the summary of tfidf model is more relevant than text rank model.

Text classification plays a significant role in information retrieval. These systems help to organize the data. N Gram text modelling technique with classifier algorithms has been used in proposed system and it has been found out that the accuracy of both the models is good with default values (unigram).

The future scope of these two systems is –

For summarization the scope can be extended to abstractive technique with addition of more NLP features.

For classification, the systems can be tested for large size corpus and new techniques or domains can be added.

6. REFERENCES

[1] S. G. Sheetal Shimpikar, "A Survey of Text Summarization Techniques for Indian Regional Languages," International Journal of Computer Applications, vol. 165, pp. 29-33, May 2017.

[2] D. M. a. D. K. Virat V. Giri, "A Survey of Automatic Text Summarization System for Different Regional Language in India," Bonfring International Journal of Software Engineering and Soft Computing,, vol. 6, pp. 52-57, October 2016.

[3] S. B. K. V. M. K. Apurva D. Dhawale, "Automatic Preprocessing of Marathi Text for Summarization," International Journal of Engineering and Advanced Technology (IJEAT), vol. 10, no. 1, pp. 230-234, October 2020.

[4] T. J. S. Mudassar M. Majgaonker, "Discovering suffixes: A Case Study for Marathi Language," International Journal on Computer Science and Engineering, vol. 2, no. 8, pp. 2716-2720, 2010.

[5] D. S. a. C. N. M. Deepali K. Gaikwad, "Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer," IOSR Journal of Computer Engineering (IOSR-JCE), pp. 51-54.

[6] M. D. J. M. V. B. P. A. D. Mr. Shubham Bhosale, "Marathi e-Newspaper Text Summarization Using Automatic Keyword Extraction Technique," International Journal of Advance Engineering and Research Development, vol. 5, no. 3, pp. 789-792, March 2018.

[7] M. P. B. G. Ms. Jayshri Arjun Patil, "Review of Name Entity Recognition in Marathi Language," IJSART, vol. 2, no. 6, pp. 497-499, June 2016.

[8] V. V. Sarwadnya, "Marathi Extractive Text Summarizer using Graph Based Model," IEEE, 2018.

[9] P. G. M. D. Nutan B. Zungre, "Sense Disambiguation For Marathi Language Words Using Graph Based Model," IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare , 2016.

[10] A. D. D. K. Anishka Chaudhari, "Marathi text summarization using neural networks," International Journal of Advance Research and Development, vol. 4, no. 11, pp. 1-3, 2019.

[11] S. G. Pooja Bolaj, "Text Classification for Marathi Documents using Supervised Learning Methods," International Journal of Computer Applications, vol. 155, no. 8, pp. 6-10, December 2016.