Multi-Document Text Summarization of Marathi Regional Language

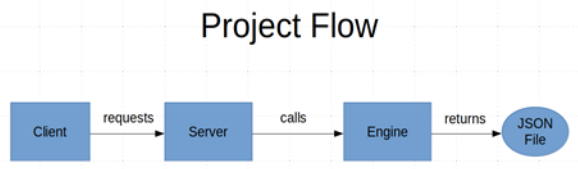
Sujata Lungare[1], Aman Jain[2], Tejashri Bhingare[3], Priyanka Tak[4], Mr.Pratik Kamble[5]
[1,2,3,4,5]*Department of IT, Marathwada Mitra Mandal's College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra*

Abstract- **In this new aeon, where enormous information is available on the internet, the best significance is to provide improved mechanism to extract information in a short period of time and efficiently. It is inconvenient for human beings to manually extract summary of large documents of text. There are ample text materials available on the internet, so it is not easy to find relevant documents from a number of documents which are currently available .In order to solve this problem Automated Text Summarization comes into picture. Text Summarization is a process of pinpointing the most important meaningful information in a document or a set of documents compressing them in to a shorter version conserving its overall meanings. In this paper the author has given a brief view of how a user will be using a web application to summarize Marathi articles or documents. Such summary is generated using probabilistic model for multi-source Marathi text summarization. Here, precision, recall, compression ratio as well as retention ratio are the parameters which are used for evaluation purpose .Implemented system successfully generates relevant summary of Marathi articles.**

Index Terms- Extractive text summarization, Text mining, Stop word removal, Bag of words.

I. INTRODUCTION

Text summarization is the process of automatically creating a flatten version of a given text which provides valuable and useful information for the user. This paper is focused on summarization of multiple Marathi documents .Our approach mainly focuses on extracting information from various newspapers which is a tedious task for each individual. As a part of a regional language, there are multiple popular Marathi e-newspapers which are made available on the internet on a free basis. Some of them are Sakal, Lokmat, Maharashtra Times etc. Talking about the summarization, it can be bifurcated into two types such as abstractive and extractive .This paper deals with Extractive Text Summarization which makes use of Machine learning based algorithms, Natural language processing (NLP), Java machine learning libraries. Extractive summarization has a property stating that all the important sentences are identified and only those sentences are included in a summary and this desired summaries length can be obtained with the help of compression ratio. With the help of NLP multiple documents are merged together, data cleaning of a particular dataset is carried out, feature selection is performed and a text model is generated. In broad terms we can say that text summarization is a process of compressing a text document in order to create a summary by jotting down the major points of a document. Automated text summarization not only saves time but also helps in reducing the document content. Text summarization tool is currently available for English language. This paper clearly highlights a presence of Marathi regional language which is carried out with the help of algorithms.



This paper is organized as follows : Section 2 explains the algorithms, Section 3 explains the Architecture, Section 4 states the algorithms used, Section 5 shows implementation and evaluation parameters, Section 6 shows the result, Section 7 concludes this paper.

II. LITERATURE SURVEY

In this section we are discussing various approaches for Stop word removal, extractive summarization and algorithms as well.

In [1], the author Irena Spasic has proposed a technology related to multi-word term recognition in which a flexiterm is extended to recognize acronyms which can be consolidated into term merging process. Not only the author has stated about the acronym recognition but has also discussed about its integration with other types of term variation. Flexi term also implements unsupervised approach of extracting multiple word terms which is carried out with the help of lexico-filtering.

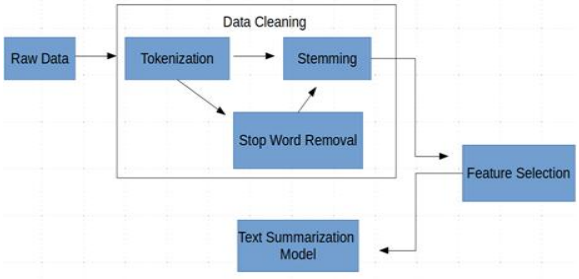
In [2], the author Yogeshwari V. Rathod has explained about the text -rank graph based ranking model which is used for graph extraction from natural language texts .Page rank algorithm is also having a leading role in this paper which consist of key phrase extraction. These consist of various modules related to pre-processing as well as stemming which are resulting in stop word removal and tokenization methods.

In [3], authors Sandeep Sripada, Venu Gopal Kasturi, Gautam Kumar Parai discussed about multi-document approach towards summarization focusing on a extractive type including a novel graph based formulation . K-means clustering algorithm is used in clustering together similar sentences from multiple documents. Stack decoder algorithm is also used by the authors in order to generate summaries close to global optimal as this algorithm is able to test multiple summary lengths.

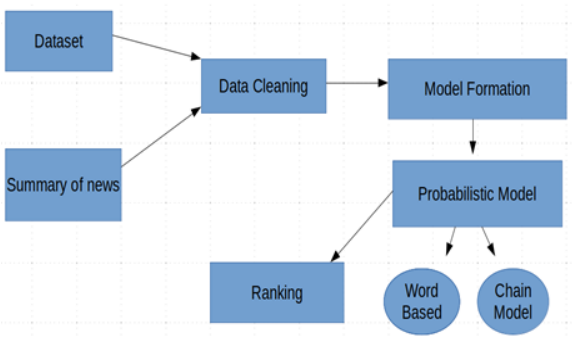
In [4], authors Shubham Bhosale, Diksha Joshi, Vrushali Bhise, Rushali A.Deshmukh discuss about the keyword extraction algorithm which works efficiently and automatically in extracting keywords in e-newspaper articles. A word extraction module is used by the user in providing the text .This statistical approach is used due to its better performance and less complexity.

III. PROPOSED ARCHITECTURE

System Architecture



TRAINING ARCHITECTURE



IV. ALGORITHM

In this section we describe two algorithms which are based on extractive text summarization.

Probabilistic Algorithm

The main focus in probabilistic algorithm is towards the probability of words that are used. This algorithm has a speciality by which multiple documents can be linked or merged together. Summarization helps in avoiding repetitions. Size of the summary is dynamic and can be automatically learned by the engine i.e the algorithm learns the size of the summary with the help of training, E.g: 100 sentences in 10 words cannot be possible .Result is totally dependent on chance. Another name for probabilistic is randomized algorithm.

One of the best example for probabilistic algorithm is the Markov Chain model which is used in selecting the sentences from a particular piece of article as well as it also helps in finding most probabilistic words in a particular sentence. Markov Chain can explain most complicated real time processes and it is said to be memory less.

Probabilistic algorithm includes a model named as Bag of Words Model (BOW) .BOW is a word parser which helps in parsing each and every sentence.

This model is simple in terms of its flexibility as well as implementation .This model states all the words which are known in a particular document. It has nothing to do where the word is located in a document.

Main attraction of BOW:

- Data collection
- Vocabulary
- Vector values
- Frequency & count

TF-IDF:-

[Term frequency & Inverse Document frequency]

It is a vector based model. This TF-IDF is a product of 2 weights, one is the term frequency and the inverse document frequency.

TF-IDF is used in multi-document as it is good in removing words which are not of much importance.

Term frequency

$$\sum_{i=0}^{n} = \frac{a_i}{Wn}$$

Where,

Term frequency is a weight representing of how often a word occurs in a document

Document frequency

$$\sum_{i=0}^{n} = \frac{O_i}{n}$$

Where,

Inverse document frequency is another weight representation of how common is a word across a document

Probability Calculation:

Probability of word (Wn)

$$\sum_{w=1}^{n} = \frac{O_w}{n}$$

2.2 Stemmers for Marathi

Stemmers are a fundamental part of natural language processing as well as information retrieval applications. Stemmers are basically used for regional languages .In this paper our approach is towards summarization of Marathi regional language, so we have used stemmers for marathi. In case of Marathi, lamitization is absent .This algorithm is mainly used in removing suffixes by using a list of frequent suffixes.

There are some modules which are based on Stemmers:

1. Pre-processing Modules
2. Stemming Modules

In Marathi language, a single root word such as महाराष्ट्रा can have assorted relations like महाराष्ट्राचा, महाराष्ट्राची, महाराष्ट्रामध्ये, महाराष्ट्रासाठी, महाराष्ट्रावर, महाराष्ट्राकड

IMPLEMENTATION

While training the input data datasets of news from various news sources is taken along with their ideal summary.

Likewise, testing of input training model with probabilities of sentences and words is performed.

Training Phase:

1. Read and parse the input.
2. Separate heading, Place, Source, Contents and Summary.
3. Process contents and calculate words, sentence frequency.
4. Map sentences from summary and contents to form probabilities.
5. Store the generated model.

Similarly testing of input data is carried out by taking datasets of news from various news sources.

The testing output shows the actual output in the form of a summary for the given set of input news data.

Testing Phase:

1. Read and parse the input.
2. Separate heading, Place, Source, Contents and Summary.
3. Process contents and calculate words , sentence frequency.
4. Apply the trained model for generating the summary based on probabilities.
5. Store summary.

EVALUATION PARAMETERS:

1. Precision(P) : This is one of the main evaluation parameters .Precision(P) is the number of sentences that occur in both system as well as in ideal summaries which is divided by the number of sentences in the system summary .

2. Recall(R): It measures the number of sentences that occur in both the system as well as in ideal summaries which is divided by the number of sentences present in the ideal summary.

3. F-Score (F): It combines both Precision (P) as well as Recall(R)

Formula for the same is listed below:

$F=2 \cdot P \cdot R / P+R$

Precision (P) = 8/10 = 80%

Recall(R) = 8/10 = 80%

VI. RESULT

Ideal Input:


Title: पुणे: बेकरीला लागलेल्या आगीत ६ कामगारांचा होरपळून मृत्यू
Source: saamana
Place: पुणे
Contents : पहाटेच्या साखर झोपेत असताना बेकरीत झोपलेल्या सहा कामगारांची ती काळरात्र ठरली.
शॉर्ट सर्किटने दुकानात आग लागल्यानंतर पोटमाळ्यावर झोपलेल्या कामगारांना जाग आली, खाली उतरून बाहेर पडण्यासाठी धडपड सुरू केली. आरडाओरड केली, पण मालकाने बाहेरून कुलुप लावल्याने त्यांचे सगळे प्रयत्न व्यर्थच जात होते. दुकानात खाली आग, पोटमाळ्यावर सगळीकडे धूर अशा परिस्थितीमध्ये अडकलेल्या कामगारांनी तडफडून एकमेकांना मिठी मारून प्राण सोडले. ही हृदयद्रावक घटना कोंढव्यात घडली. इरशाद खान (वय २६), शान अन्सारी (वय २२), जाकीर अन्सारी (वय २४), फहिम अन्सारी (वय २१), जुनेद अन्सारी (वय २५), निशान अन्सारी (वय २१, सगळे मूळ रा बिजनौर, उत्तरप्रदेश) अशी मृतांची नावे आहेत. अब्दुल मोहम्मद युसुफ चिन्नीवार (वय २७, कुमार होमस, एनआयबीएम रस्ता कोंढवा) असे बेकरी मालकाचे नाव आहे. कोंढव्यात ताताब मशिदीसमोर गगन अव्हेन्यूव नावाच्या सोसायटीत "बेकस अॅन्ड केकस" ही बेकरी आहे. तेथील कामगार दिवसभर काम करून तेथेच झोपतात. रोजच्याप्रमाणे काम संपल्यानंतर मालकाने बेकरीला बाहेरून कुलुप लावले. त्यानंतर सहा कामगार पोटमाळ्यावर जाऊन झोपले. पहाटे सवा चारच्या सुमारास बेकरीत शॉर्ट सर्किटने आग लागली, त्यानंतर कामगारांना जाग आली. त्यांनी बाहेर पडण्यासाठी खटाटोप सुरू केला. पण बाहेर पडता येत नव्हते. द कानातून धूर येताच नागरिकांनी अग्निशमन दलाला खबरी दिली. अग्निशमन दलाचे पथक, रूग्णवाहिका घटनास्थळी पोहोचले. बेकरीचे कुलुप तोडून शटर उघडून पाण्याचा मारा करून आग आटोक्यात आणण्यात आली, दरम्यान पोटमाळ्यावर गेलयानंतर तेथे सहा जण एकमेकांना घट्ट पकडून बेशुद्ध अवस्थेत पडलयाचे धक्कादायक दृश्य अग्निशमन दलाच्या जवानांनी पाहिले. कामगारांची वैद्यकीय तपासणी केलयावर त्यांचा मृत्यू झालयाचे लक्षात आले. पोलिसांनी या प्रकरणी पंचनामा करून गुन्हा दाखल केला आहे.

Ideal Output:


शॉर्टसर्किटमुळे बेकरीला लागलेल्या आगीमध्ये सहा कामगारांचा झोपेतच होरपळून मृत्यू झाल्याची भीषण घटना समोर आली आहे. एप्रिल 2015 मध्ये ही बेकरी सुरू करण्यात आलेली आहे. अग्निशामक दलाच्या अधिकाऱ्यांनी दिलेल्या माहितीनुसार, कोंढव्यातील गगन एव्हेन्यू इमारतीमध्ये 'बेकस अॅन्ड केकस' नावाची बेकरी आहे. आकाराने छोट्या असलेल्या या बेकरीच्या पोटमाळ्यावर हे कामगार झोपलेले होते. तर पोटमाळ्यावर पीठ मळण्यासाठी मोठर बसवण्यात आलेली आहे. पहाटे सवा चारच्या सुमारास बेकरीत शॉर्ट सर्किटने आग लागली, त्यानंतर कामगारांना जाग आली त्यावेळी एकमेकांना बिलगून झोपलेल्या अवस्थेतच सहा कामगारांचे मृतदेह पडलेले होते. मृतांमध्ये 20 ते 26 या वयोगटातील तरुणांचा समावेश आहे हृदयद्रावक आणि भयावह असे ते चित्र होते. बेकरीला बाहेरून कुलुप लावण्यात आले असल्याने कामगारांना बाहेर निघता आले नाही. हे कामगार जर खालच्या भागात झोपले असते तर कदाचित त्यांचे प्राण वाचू शकले असते अग्निशमन दलाच्या जवानांनी घटनास्थळी पोहोचून ही आग विझवली आणि कामगारांचे मृतदेह बाहेर काढले. बेकरीला बाहेरून कुलुप लावण्यात आले असल्याने कामगारांना बाहेर निघता आले नाही. हे कामगार जर खालच्या भागात झोपले असते तर कदाचित त्यांचे प्राण वाचू शकले असते अग्निशमन दलाच्या जवानांनी घटनास्थळी पोहोचले. बेकरीचे कुलुप तोडून शटर उघडून पाण्याचा मारा करून आग आटोक्यात आणण्यात आली, दरम्यान पोटमाळ्यावर गेलयानंतर तेथे सहा जण एकमेकांना घट्ट पकडून बेशुद्ध अवस्थेत पडलयाचे धक्कादायक दृश्य अग्निशमन दलाच्या जवानांनी पाहिले. कामगारांची वैद्यकीय तपासणी केलयावर त्यांचा मृत्यू झालयाचे लक्षात आले. पोलिसांनी या प्रकरणी पंचनामा करून गुन्हा दाखल केला आहे.

Input:


Title: पुण्यात बेकरीला लागलेल्या आगीत सहा कामगारांचा मृत्यू
Source: loksatta
Place: पुणे
Contents :
पुण्यातील कोंढवा बद्रुक्मधील बेकरीला लागलेल्या आगीत सहा कामगारांचा मृत्यू झाला आहे. ताताब कंपनीजवळ असणाऱ्या बेकस एँण्ड केकस बेकरीला पहाटेच्या सुमारास आग लागली होती. या बेकरीमध्ये सहा कामगार झोपेते होते. या कामगारांचा आगीत होरपळल्याने मृत्यू झाला. अग्निशमन दलाच्या जवानांनी घटनास्थळी पोहोचून ही आग विझवली आणि कामगारांचे मृतदेह बाहेर काढले. बेकरीला बाहेरून कुलुप लावण्यात आले असल्याने कामगारांना बाहेर निघता आले नाही. शॉर्ट सर्किटमुळे ही आग लागल्याचा प्राथमिक अंदाज वर्तवला जातो आहे. अग्निशमन दलाच्या जवानांनी या आगीवर संपूर्ण नियंत्रण मिळवले असून कलिम ऑपरेशनदरम्यान पुरण केले आहे. बेकरीला लागलेल्या आगीत मृत्युमुखी पडलेले सगळे कामगार उत्तर प्रदेशचे रहिवासी आहेत. इरशाद खान, शान अन्सारी, झाकीर अन्सारी, फहिम अन्सारी, जुनेद अन्सारी, निशाण अन्सारी अशी मृत्युमुखी पडलेल्या कामगारांची नावे आहेत.

Output:


शॉर्टसर्किटमुळे बेकरीला लागलेल्या आगीमध्ये सहा कामगारांचा झोपेतच होरपळून मृत्यू झाल्याची भीषण घटना समोर आली आहे. आगीमध्ये जवळपास निम्मी बेकरी जळून खाक झाली. एप्रिल 2015 मध्ये ही बेकरी सुरू करण्यात आलेली आहे. 'बेकस अॅन्ड केकस' ही बेकरी अब्दुलला मोहम्मद युसुफ चिन्नीवार (वय २७, रा कुमार होमस, एनआयबीएम रस्ता, कोंढवा), तयब अन्सारी (वय २६, रा सय्यदनगर, हडपसर), मुनीर चिन्नीवार (वय ६२, रा पारगे नगर, कोंढवा) यांच्या भागीदारी मालकीची आहे आकाराने छोट्या असलेल्या या बेकरीच्या पोटमाळ्यावर हे कामगार झोपलेले होते पहाटे सवा चारच्या सुमारास बेकरीत शॉर्ट सर्किटने आग लागली, त्यानंतर कामगारांना जाग आली. त्यावेळी एकमेकांना बिलगून झोपलेल्या अवस्थेतच सहा कामगारांचे मृतदेह पडलेले होते. हे कामगार जर खालच्या भागात झोपले असते तर कदाचित त्यांचे प्राण वाचू शकले असते. मृतांमध्ये 20 ते 26 या वयोगटातील तरुणांचा समावेश आहे.

VII. CONCLUSION

In this paper Text summarization system using Marathi news articles is discussed and implemented. System uses extractive machine learning algorithm involving probabilistic approach. Summarization of multiple marathi documents together in a single summary can be useful for various marathi readers across the globe. Generated summary is seen to have extracted relevant and important contents as desired. Currently, system is implemented as a web service. In future, it can be further extended to various regional languages like Hindi. Also, it can be pushed on various platforms like mobile apps or cloud for greater reach.

REFERENCES

[1] Irena Spasic , Acronyms as an Integral Part of Multi-Word Term Recognition - A Token of Appreciation, School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K

[2] Yogeshwari Rathod, Extractive Text Summarization of Marathi News Articles, Department of Computer Science & Engineering, Vishwakarma Institute of Technology , Pune

[3] Sandeep Sripada , Venu Gopal Kasturi , Gautam Kumar Parai ,Multi-Document extraction based Summarization

[4] Shubham Bhosale, Diksha Joshi , Vrushali Bhise, Rushali. A.Deshmukh , Marathi e-Newspaper Text Summarization Using Automatic Keyword Extraction Technique , Volume 5,Issue 03 ,March -2018 , Rajashri Shahu College of Engineering , Pune

[5] Giines Erkan , Dragomir R.Radev , Lex PageRank : Prestige in Multi-Document Text Summarization , Department of EECS , School of Information Technology , University of Michigan

[6] H.Li , J.Zhu , Modal Summarization for Asynchronous Collection of Text , Conference of EMNLP , 2017

[7] Josef Steinberger , Karel Jezek , Evaluation Parameters for Text Summarization , Department of Computer Science and Engineering , University of West Bohemia in Pilsen , Czech Republic

[8] NFPA 1852: Standard on Selection, Care, and Maintenance of Open-Circuit Self-Contained Breathing Apparatus (SCBA), 2019 Edition. In NFPA National Fire Codes Online. Retrieved from http://codesonline.nfpa.org (In association with IISc Bangalore,Karnataka)