International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.8 January - February – 2024
DOI:10.46647/ijetms.2024.v08i01.026 ISSN: 2581-4621

Key phrase extraction from patient's chief complaints

Bhanudas Suresh Panchbhai[1], Dr.Varsha Makarand Pathak[2]

[1]*Department of computer science, R.C.Patel Arts Commerce and Science College, Shirpur, Maharashtra, India*
[2]*Department of computer application, KCES'S Institute of Management and Research,Jalgaon, Maharashtra, India*

ABSTRACT
Over the past ten years, there have been notable advancements in natural language processing(NLP) systems. The goal of the study is to extract relevant key phrases from the Marathi patient major complaints. Our data came from a single, non-profit private hospital located in a remote part of North Maharashtra Region, India. From the hospital's information around 5000 patient visits occurred between 2021 and 2022.A total of 54000 text-free chief complaints were gathered as a result of the initiatives. This paper summarizes a study that focuses on pre-processing strategies that were established in an efficient way to extract pertinent keywords from phrases in accordance with the patients' primary concerns. The methods and findings from this experiment are presented in this paper.
Keywords: Key terms, Extraction, Marathi, Pre-processing, Natural languageprocessing.

- **Problem Statement**

• This work, which is presented here, focuses on text pre-processing algorithms that have been created to effectively extract pertinent key terms from phrases that correspond to patients' chief complaints
• Mapping of those key phrases to relevant medical terminology. The technique and findings of this study's experimental efforts are presented.

- Challenges faced

• Lack of availability of tools to develop Marathi domain ontology in medical health care.
• Data Collection of proper patient's chief complaints from various hospitals from North Maharashtra Region, India

1. Introduction

Text pre-processing is an essential and fundamental stage in Natural Language Processing before developing any model. In order to clear up ambiguity and inconsistencies from a raw text corpus that has been gathered from one or more sources, pre-processing is required [1].In order to access precise and dependable data, a user must devise the most effective technique available. Even after cleaning, more text pre-processing is required to rearrange the data so that it may be immediately entered into the model. If text pre-processing is not done properly, the data will be as useful as useless, and the NLP model produced will only be as bad as useless. Many well-known languages have been the subject of research in the field of text processing. However, regional language text is now very necessary. In Maharashtra, where the Marathi language is somewhat less focused, this has been taken into account when analyzing the work for regional languages. This study uses text pre-processing techniques to extract pertinent key phrases from phrases and a mapping of the key phrases with medical terminology that correspond to the patients' primary concerns. This study was created to extract pertinent key phrases from phrases in accordance with the patients' main concerns and it relies on text pre-processing approaches. The article that follows covers the methodology, findings, and literature review.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.8 January - February – 2024
DOI:10.46647/ijetms.2024.v08i01.026 ISSN: 2581-4621

2. Literature Survey

Sr.No.	Author Name & References	Aim
1	Afham Fardeen[1] : Techniques of Text Preprocessing Using NLTK inPython.	In order to clear up ambiguity and inconsistencies from a raw text corpus that has been gathered from one or more sources, pre-processing is required.
2	S.P. Paramesh, K.S. Shreedhara [2]: IT Help DeskIncident Classification UsingClassifier Ensembles.	Claims that in order to achieve better results, the description data for IT tickets needs to be cleaned up. The author employed POS tagging, stemming, and a list of frequent English stop words to eliminate keywords from the text.
3	Hawari and Hala Barham [3]:A machine learning based help desk system for IT service management.	Examined the prediction accuracy of classification models by training them on two different datasets, one that had been preprocessed and the other that hadn't. The experiment demonstrated that each of the four classification models had an accuracy improvement of about 20 to 30 percent using preprocessed and cleaned description data.
4	Paramesh S.P, Shreedhara K.S [4]: Building Intelligent Service Desk Systems using	There has been a focus on efficient text preparation. Regular expressions, useful for pattern matching, are used to remove stop words along with date, time, and numeric data. To
	AI.	balance the uneven data, the author has employed random oversampling and under sampling approaches. Feature selection follows feature extraction.
5	.N. Vasunthira Devi [5]: A Systematic Survey of Natural Language Processing (NLP) Approaches in Different Systems.	The limitations and usefulness of NLP techniques employed in healthcare and education were examined. Effective grammar and corpora can be used by educational institutions to aid students in the development of their talents.
6	Judith D. Trippe [6]: A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques	Shown how text pretreatment has a major impact on every process in the final stages of natural language processing. Stemming, filtering, and tokenization are techniques the author employs.
7	Virat V. Giri, Dr. U.P. Kulkarni, and Dr. M.M. Math[7]: Survey on Pre- Processing Techniques forText Mining.	The concise summary given by the summarizing approach allows readers to swiftly and readily understand the content of original papers without having to read each one individually.

DOI:10.46647/ijetms.2024.v08i01.026 ISSN: 2581-4621

8	Ranjan Satapathy, Erik Cambria[8] :Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis	Has released two models for the micro text language most commonly used on social networking platforms like Twitter.
9	Virat V. Giri, Dr.M.M. Math& Dr.U.P. Kulkarni[9]: A Survey of Automatic Text Summarization System for Different Regional Languagein India.	Automatic text summarization is a method for cutting the original text's length while retaining all of its content and meaning.
10	Hovy, E., & Lin, C. Y[10]: Automated text summarization and theSUMMARIST system	This essay is divided into three sections: a preliminary typology of summaries in general; a description of the SUMMARIST automatic multilingual text summarization system currently being developed at ISI; and a discussion of three approaches for evaluating summaries.
11	Jovi D'silva, Dr.Uzzal Sharma [11]: Text Summarization using Rich Semantic Graph for MarathiSentence.	It is possible to condense a lengthy text into a shorter version using a process called text summarizing without losing any of the text's original meaning.
12	Sheetal Shimpikar, Sharvari Govilkar[12]: Automatic Text Summarization of Indian Languages: A Multilingual Problem.	To summarize news that will be useful to students taking competitive tests, we are concentrating on educational, political, and sports news.

Many automatic text pre-processing techniques are available for most commonly used natural languages. Only English and other international languages are frequently supported by these text pre-processing methods. There are fewer automatic text pre-processing techniques available for Indian languages. The following section discusses various NLP-based text pre- processing strategies for Indian languages:

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.8 January - February – 2024
DOI:10.46647/ijetms.2024.v08i01.026 ISSN: 2581-4621

3. Proposed Architecture

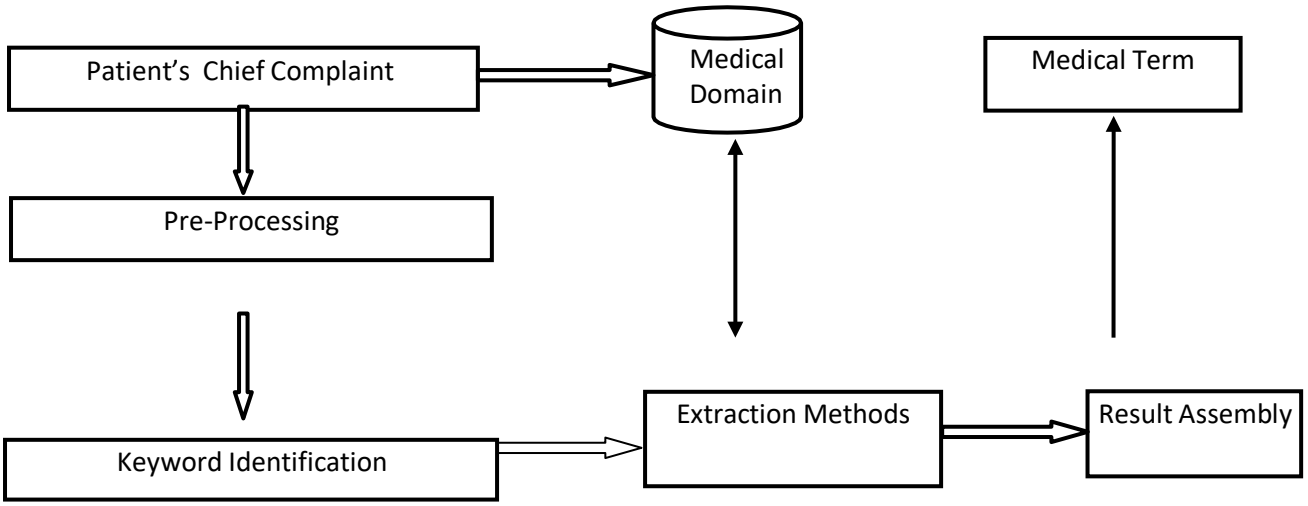

Fig.1. Architecture of Proposed Work

4. Materials and Methods

1. Data Collection

- **Setting** A single, private, and nonprofit hospital in rural (Dhule Dist.), NorthMaharashtra Region, and India.
- **Participants** From 2020 to 2021, 3600 patient visits produced 4000 total textfree chief complaints in Marathi language used for the validation dataset.
- In this study, we take a sample of 150 complaints out of 4000 total complaintsin .csv.
- In this study, we make use of a patient symptoms database, which is 82.5 KBin size.

2. Algorithm 1: Pre-processing
An algorithm to extract and preprocess the number of important phrases from patientchief complaints 150 sample patient chief complaints in Marathi were entered.Input: 82.5 KB, Marathi-language symptom database
Steps:
1. Open & read patient chief complaints file.
2. For each sentence apply pre-processing:Use try:
- Calculate length of sentence.
- Apply tokenization.
- Filtered the text.
- Find frequency of each word in text.
- Apply stemming to tokens.
- Apply lemmatization to tokens.
- Find Pos for text tokens.
- Perform chunking.
- Append these tokens to one list valid dataset of chief complaints.Exceptions
- Print exceptions

3. Open & read both files one as a symptom database file and second is valid datasetof chief complaints.
- Presentkwl = set (symptom database)
- For each sentence in valid chief complaints:
- For each word in sentence:

If word in symptom database

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.8 January - February – 2024
DOI:10.46647/ijetms.2024.v08i01.026 ISSN: 2581-4621

- Presentkwl.add (word)Else
- Filteredlist.append (word)

4. Print Presentkwl in .csv format.
5. Stop.

Implementation of above algorithm by using Pandas Python code

1. import pandas as pd
2. import re
3. file_path = "Patient Database2510.xlsx"
4. try:
5. # Read the Excel file into a pandas DataFrame
6. data_frame = pd.read_excel(file_path)
7. # Extract the 'Chief_complain' column data as a list
8. chief_complain_data = data_frame['Chief_complain'].tolist()
9. chief_complain_set = set(chief_complain_data)
10. # print(chief_complain_set)
11. # print(chief_complain_data)
12. # with open(file_path, "r", encoding="utf-8") as file:
13. # data_as_string = file.read()
14. # keywordsSet = set(data_as_string.split())
15. # print(keywordsSet)
16. except Exception as e:
17. print("Error reading the file: {}".format(e))18. pattern = r'[.,\'\"!]'
19. string = '[illegible].'
20. string = re.sub(pattern, '', string)
21. wordsList = string.split()
22. print('WordsList: ',wordsList)
23. filteredList = []
24. presentKWL = set()

DOI:10.46647/ijetms.2024.v08i01.026 ISSN: 2581-4621

```
for word in wordsList:
if word in chief_complain_set:
presentKWL.add(word)
else:
filteredList.append(word)
print('Keywords Present: ', presentKWL)
print('Filtered List: ',filteredList)
```

```
32. updatedString = " ".join(filteredList)
33. print(updatedString)
```

After processing data with the toolkit in Python, we receive the output as key terms.

5. Results

Sr.No.	Technique	Complaints	Precision	Recall	F1 Score
1.	Key Phrase Extraction	150	0.80	0.84	0.81

6. Conclusions

It is necessary to concentrate key terms extraction on the regional language. In this study, Marathi, the native tongue of Maharashtra, is highlighted. The algorithm used to pre-process the Marathi text was effective since it changed depending on the language and pre-processing methods used. The study emphasizes the pre-processing carried out on the Marathi text to extract important phrases. First, the sample input file for patient complaints is opened and read. We then pre-process text to create legitimate tokens. We then compare the pattern of each sentence's valid tokens with databases of symptoms to extract key terms. We are developing a system that is considerably more powerful and efficient in order to extract medical key terms from Marathi text**.**

7. References

[1] Afham Fardeen,Techniques of Text Preprocessing Using NLTK in Python ,June 20, 2021.
[2] S.P. Paramesh, K.S. Shreedhara," IT Help Desk Incident Classification Using Classifier Ensembles", ICTACT Journal On Soft Computing, July 2019, Vol: 09, Issue: 04.
[3] Feras Al-Hawari, Hala Barham, "A machine learning based help desk system for IT service management", Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.04.001.
[4] Paramesh S.P, Shreedhara K.S,"Building Intelligent Service Desk Systems using AI", IJESM 2019; Vol: 1, No: 2, pp: 01-08.
[5] N.Vasunthira Devi, Dr. R. Ponnusamy, "A Systematic Survey of Natural Language Processing (NLP) Approaches in Different Systems,",IJCSE 2016, vol. 4,issue 7, pg 192-198.
[6] Elizabeth D. Trippe , Krys Kochut , Juan B. Gutierrez, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919v2 [cs.CL] 28 Jul 2017.
[7] Dr. Balasubramani, Arjun Srinivas Nayak, Ananthu P Kanive, Naveen Chandavekar,"Survey on Pre-Processing Techniques for Text Mining" ,IJECS, Volume 5 Issue 6 June 2016, Page No. 16875-16879.
[8] Ranjan Satapathy, Claudia Guerreiro, Iti Chaturvedi, Erik Cambria,"Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis", 2017 IEEE International Conference on Data Mining Workshops.
[9] Virat V. Giri, Dr.M.M. Math & Dr.U.P. Kulkarni, "A Survey of Automatic Text Summarization System for Different Regional Language in India", In Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6, Special Issue, October 2016.
[10] Hovy, E., & Lin, C. Y. (1998, October). Automated text summarization and the SUMMARIST system. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998 (pp. 197-214). Association for Computational Linguistics.
[11] Sheetal Shimpikar, Sharvari Govilkar, "Abstractive Text Summarization using Rich Semantic Graph for Marathi Sentence", JASC: Journal of Applied Science and Computations Volume V, Issue XII, ISSN NO: 1076-5131, December/2018.
[12] Jovi D'silva, Dr.Uzzal Sharma, "Automatic Text Summarization of Indian Languages: A Multilingual Problem", Journal of Theoretical and Applied Information Technology Vol.97. No 11, 15th June 2019.