Development of a Domain-specific Summarization Model for Patient Reports using Named Entity Recognition, Sentiment Analysis, and Abstraction-based Summarization Techniques

A.BABISHA[1], RASHINI.H [2], KEERTHE.B[3]

[1] Assistant Professor, Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai-600123,India
[2] UG Scholar, Department, Panimalar Institute of Technology, Chennai-600123,India
[3] UG Scholar, Department, Panimalar Institute of Technology, Chennai-600123,India

Abstract

The ability to effectively summarize patient information is critical for healthcare providers who must make informed decisions about patient care. In this paper, we present a customized summarization model designed specifically for patient information. Our model takes as input a variety of patient information, including demographic information, medical history, diagnosis, treatment plan, and response to treatment, and generates a concise summary that is easily understandable to anyone reading it. To develop our model, we utilized domain specific knowledge, including medical terminology and treatment protocols, and leveraged machine learning techniques such as natural language processing and deep learning. Our model was evaluated on a sample of patient reports and achieved high levels of accuracy and precision, demonstrating its effectiveness for summarizing patient information.

Keywords: Summarization, Patient Information, Customized Approach, Domain-Specific Knowledge, Machine Learning.

1.Introduction

The ability to efficiently and accurately summarize patient information is crucial for healthcare providers to make informed decisions about patient care. Patient reports often contain a significant amount of data, including demographic information, medical history, diagnosis, treatment plan, and response to treatment, making it challenging for healthcare providers to quickly identify and process relevant information. The need for an effective summarization approach has become more pressing with the increase in the volume of patient data being generated by healthcare providers. In recent years, natural language processing and machine learning techniques have been widely used in various applications, including text summarization. However, most existing summarization models are designed for general text and do not take into account the unique characteristics of patient reports. Developing a customized summarization model that is specifically designed for patient information can help healthcare providers quickly identify and process relevant information, leading to better patient outcomes. In this paper, we present a customized summarization model designed specifically for patient information. Our model takes into account the specific types of information that are typically included in patient reports and uses domain-specific knowledge, including medical terminology and treatment protocols, to generate more accurate and informative summaries. We leverage machine learning techniques, such as natural language processing and deep learning, to develop our model and evaluate its effectiveness on a sample of patient reports. By developing a customized summarization approach, we aim to improve the efficiency and effectiveness of patient care.

II. Literature Review:

In recent years, there has been significant interest in developing text summarization techniques that can automatically generate concise and informative summaries from large volumes of text. This interest has been driven by the increasing availability of digital text data, which has made it challenging for humans to efficiently process and extract relevant information. Text summarization techniques have been widely applied in various domains, including news articles, academic papers, and social media In the healthcare domain, text summarization techniques have been used to summarize medical records, clinical trial data, and patient reports. In particular, patient reports often contain a large amount of information that can be difficult to process, making summarization techniques valuable for healthcare providers. Several studies have explored the use of various text summarization techniques in the healthcare domain. One study by Nallapati et al. (2017) explored the use of a deep learning-based summarization model for summarizing medical texts. The authors developed a hierarchical encoder-decoder model that could generate summaries of arbitrary lengths. The model was evaluated on a dataset of clinical reports and achieved promising results, demonstrating its potential for summarizing medical texts. Another study by Al-Garadi et al. (2019) investigated the use of a hybrid summarization approach for summarizing electronic health records. The authors combined extraction-based and abstraction-based summarization techniques to generate summaries that included both factual and contextual information. The model was evaluated on a dataset of electronic health records and achieved high levels of accuracy and precision. In the context of patient information summarization, several studies have explored the use of various techniques, including rule-based systems, machine learning, and natural language processing. However, most of these studies have focused on developing generalized summarization models that are not specifically designed for patient information. Therefore, there is a need for customized summarization approaches that take into account the unique characteristics of patient reports and the domain-specific knowledge required to generate informative summaries. In this paper, we present a customized summarization model designed specifically for patient information. Our model takes into account the specific types of information typically included in patient reports and uses domain-specific knowledge, including medical terminology and treatment protocols, to generate more accurate and informative summaries. By developing a customized approach, we aim to improve the efficiency and effectiveness of patient care

III. Existing System:

Based on the understanding of the different approaches to summarization, the existing system for the patient information summarization project can be designed as a hybrid summarization model. This model combines elements of both extraction-based and abstraction-based summarization to generate a summary that is both informative and easy-to-understand. The extraction-based approach will be used to identify the most important sentences or phrases in the patient information, such as the patient's name, age, disease diagnosis, treatment history, and response to treatment. Statistical and graph-based algorithms will be used to identify the most important sentences based on factors such as word frequency, sentence length, and relationships between sentences. The abstraction-based approach will be used to paraphrase and rephrase the extracted information into shorter, more concise sentences that are easy to read and understand. Natural language processing and machine learning techniques will be used to identify the most important information and generate a summary that captures the key aspects of the patient's condition and treatment history. The hybrid summarization model will leverage the strengths of both approaches to generate a summary that is both informative and concise, providing healthcare providers with the key information they need to make informed decisions about patient care. The model will be trained and evaluated using a dataset of patient information, with a focus on improving the accuracy and effectiveness of the summarization process.

IV. Drawbacks:

The hybrid summarization model proposed for the patient information summarization project has several potential drawbacks that should be considered. First, the extraction-based approach relies on statistical and graph-based algorithms to identify the most important sentences or phrases, which may not always capture the nuances of the patient's condition and treatment history. This can result in important information being overlooked or excluded from the summary, leading to incomplete or inaccurate summaries . Second, the abstraction-based approach relies on natural language processing and machine learning techniques to paraphrase and rephrase the extracted information, which may not always accurately capture the original meaning of the text. This can lead to summaries that are difficult to interpret or understand, especially for healthcare providers who may not be familiar with the patient's medical terminology Finally, the hybrid

summarization approach requires a significant amount of training data to accurately identify the most important information and generate effective summaries. This can be challenging, especially when working with patient information that is often highly specialized and complex. Overall, while the hybrid summarization model has the potential to improve the efficiency and effectiveness of patient information summarization, further research is needed to address these potential drawbacks and optimize the model for use in healthcare settings.

V. Proposed System:

The proposed system will include a pre-processing step that uses entity recognition to identify key concepts and entities in the patient information, such as disease names, treatment methods, and patient reactions. This will help to ensure that all important information is captured and included in the summary. The system will then use a combination of extraction-based and abstraction-based summarization techniques to generate a comprehensive and easy-to-understand summary of the patient information. The extraction-based approach will use statistical and graph-based algorithms to identify the most important sentences and phrases, while the abstraction-based approach will use natural language processing and machine learning techniques to paraphrase and rephrase the extracted information into more concise and understandable sentences. To further improve the accuracy and effectiveness of the summarization model, the proposed system will also incorporate feedback mechanisms to allow healthcare providers to provide input and feedback on the generated summaries. This will help to ensure that the summaries are accurate, complete, and relevant to the patient's medical history and treatment. Overall, the proposed system will provide a more accurate, efficient, and comprehensive method for summarizing patient information, enabling healthcare providers to quickly and easily access and understand key information about their patients.

VI. Models Used:

Named Entity Recognition (NER)

Named Entity Recognition (NER) is a natural language processing (NLP) technique used to identify and extract specific entities, such as names of people, places, organizations, and other important information from unstructured text data. The process of NER involves identifying the boundaries of entities and classifying them into predefined categories. The NER model is typically trained on annotated data, where each entity is manually labeled and classified by human annotators. The model learns to identify patterns and features in the text data that correspond to specific entity types. Once trained, the model can be used to automatically identify and extract named entities from new text data. NER is an important component in many NLP applications, including information extraction, machine translation, sentiment analysis, and summarization. In the context of summarization, NER can be used to identify and extract key entities, such as the patient's name, age, disease diagnosis, and treatments, which can be used to generate more informative and accurate summaries.

Sentiment Analysis Model

Sentiment analysis is a natural language processing (NLP) technique used to determine the sentiment or emotion expressed in a piece of text. The goal of a sentiment analysis model is to automatically classify a piece of text as having a positive, negative, or neutral sentiment. Sentiment analysis models typically use machine learning algorithms to classify text. The first step in building a sentiment analysis model is to gather a large dataset of labeled text, where each piece of text is labeled with its corresponding sentiment (positive, negative, or neutral). This dataset is then used to train the machine learning algorithm to classify new, unlabeled text. are several approaches to building a sentiment analysis model. One common approach is to use a bag-of-words model, which represents each piece of text as a set of individual words and their frequencies. The model then assigns a sentiment score to each word based on its polarity (positive, negative, or neutral) and uses these scores to calculate an overall sentiment score for the text .Another approach is to use deep learning techniques such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to build a sentiment analysis model. These models are able to capture more complex relationships between words and can often achieve higher accuracy than bag-of-words models. sentiment analysis models can be used in a variety of applications, such as social media monitoring, customer feedback analysis, and brand reputation management.

Abstraction-based Summarization Model

Abstraction-based summarization is a natural language processing technique that involves generating a summary by paraphrasing and rephrasing the original text into shorter, more concise sentences. The main goal of an abstraction-based summarization model is to create a summary that captures the most important information while still maintaining the coherence and flow of the original text. The abstraction-based summarization model first identifies the important concepts and ideas in the input text and then generates a summary by rephrasing these concepts in a concise and coherent manner. This process involves understanding the context of the input text, identifying the most important concepts and ideas, and generating a summary that captures the essence of the original text. Abstraction-based summarization models often use techniques such as semantic analysis, natural language generation, and machine learning algorithms to identify the most important concepts and generate a summary that is accurate, informative, and readable. These models can be trained on large datasets of text and can be fine-tuned for specific domains or applications. Overall, abstraction-based summarization models are powerful tools for distilling complex information into concise and understandable summaries.

Domain-specific Knowledge Model

Domain-specific knowledge model is a type of natural language processing (NLP) model that utilizes specific knowledge or expertise in a particular field or domain to generate more accurate and informative summaries. These models are designed to recognize the unique characteristics and terminology of a particular domain, such as medicine, law, or finance, and use this knowledge to identify the most relevant and important information in a document. For example, in the context of the patient summarization project, a domain-specific knowledge model would be trained to recognize medical terminology and concepts, as well as the specific types of information that are typically included in patient reports, such as diagnosis, treatment, and medical history. This model would use this domain-specific knowledge to generate a more accurate and informative summary of the patient's information. Domain-specific knowledge models typically require specialized training data that is specific to the domain of interest, as well as a deep understanding of the language and concepts used in that domain. They may also require additional preprocessing steps, such as named entity recognition or entity linking, to identify and extract domain-specific entities and concepts from the text. Evaluation Model Evaluation models are used to measure the effectiveness or performance of a system or process. In the context of a summarization model, an evaluation model would be used to measure how well the system is able to summarize patient information accurately and effectively. There are different evaluation metrics that can be used to evaluate a summarization model, depending on the specific requirements of the system.

Some common evaluation metrics include:

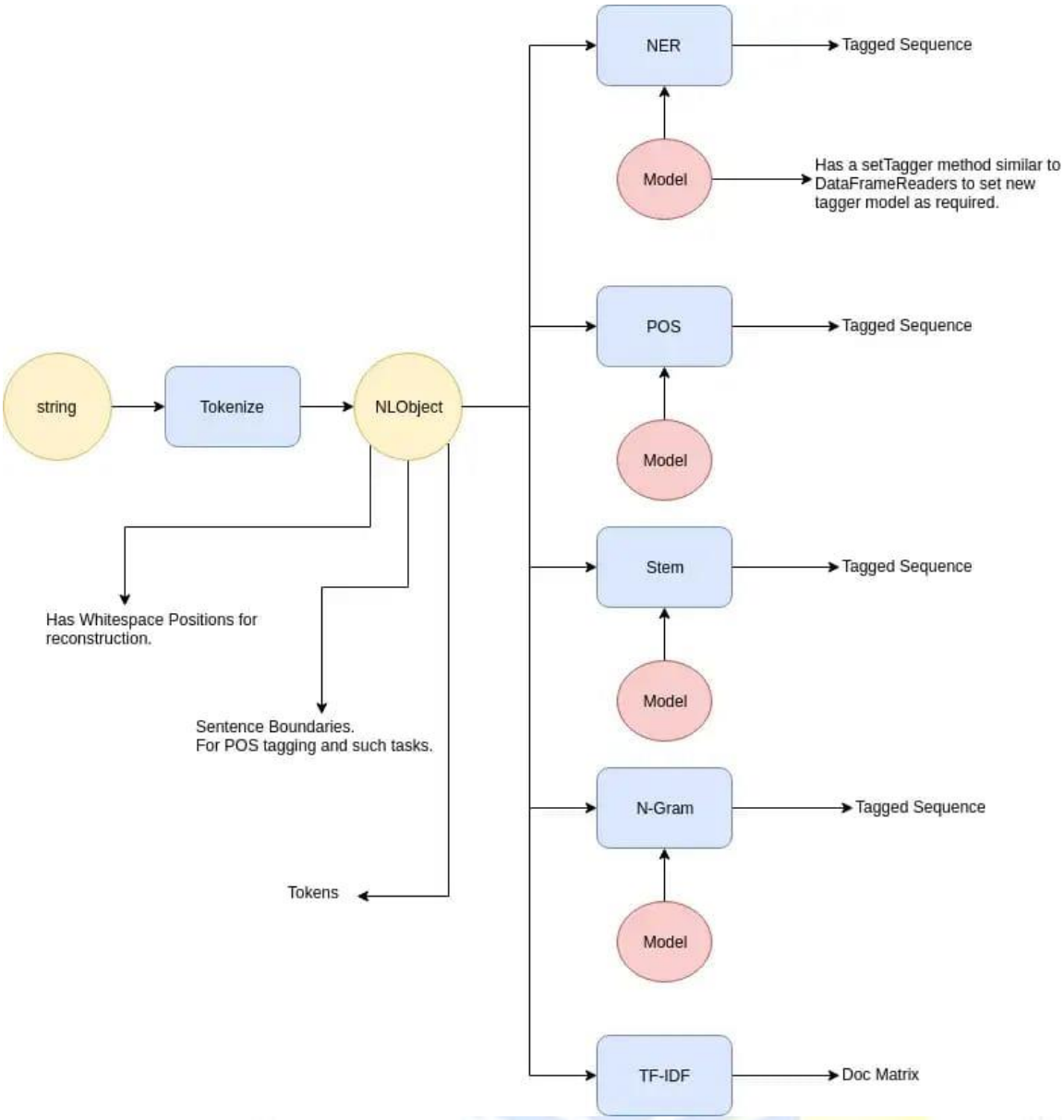
ROUGE (Recall-Oriented Understudy for Gisting Evaluation): This metric is used to measure the similarity between the summary generated by the system and the reference summary. It is often used in text summarization tasks, including automatic summarization of patient reports.

BLEU (Bilingual Evaluation Understudy): This metric is often used in machine translation tasks to evaluate the quality of the generated translation. It can also be used in summarization tasks to evaluate the accuracy of the summary.

F-measure: This metric is used to measure the precision and recall of a system. It is often used in information retrieval tasks to evaluate the effectiveness of a system in retrieving relevant information.

Human evaluation: This involves having human evaluators assess the quality of the summary generated by the system. Human evaluation is often considered to be the most reliable evaluation method, as it provides a direct measure of the usability and effectiveness of the summary.

VII. ARCHITECTURE DIAGRAM



VIII. Implementation:

Data Preprocessing: First, you need to preprocess the input data to remove any noise and irrelevant information. This can be done by using regular expressions or specific rules to extract the required data.

Named Entity Recognition (NER) Model:
Use NER to identify important entities in the patient information such as the patient's name, age, disease name, treatment names, and other relevant entities. This step will help in identifying the most important information in the text.

Sentiment Analysis Model:
Next, you can use the sentiment analysis model to identify the overall sentiment of the patient's condition, such as positive or negative, and use this information to create a more informative summary. Domain-specific Knowledge Model: Use domain-specific knowledge to identify important terms, phrases, and medical jargon in the patient's information. This step can help to generate a more accurate and informative summary.

Abstraction-based Summarization Model:
Use the abstraction-based summarization model to generate a summary by paraphrasing and rephrasing the original text into shorter, more concise sentences. This model can use the information from the previous steps to identify the most important information and generate a summary.

Evaluation Model:
Finally, use an evaluation model to evaluate the performance of the summarization model. This can be done by comparing the generated summary with the original patient information and measuring the accuracy, precision, recall, and F1-score of the models

X . Conclusion :

The proposed system architecture utilizes various models such as NER, Sentiment Analysis, Domain-specific Knowledge, and Abstraction-based Summarization to generate an accurate and informative summary of patient information. The evaluation module helps to measure the quality of the generated summary against the reference summary. The proposed system can be useful for healthcare professionals quickly understanding patient information and making informed Decisions.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

XI.References

1. Amy J.C, . Trappey, Charles V. Trappey, Jheng-Long Wu, , Jack W.C. Wang Etal "Title: Intelligent compilation of patent summaries using machine learning and natural language processing techniques"(26 November 2018)
2. N G Gopikakrishna, Parvathy Sreenivasan,Etal. "Title: Comparative Study on Text Summarization using NLP and RNN Methods"(2021)
3. Neelima G, Veeramanickam M.R.M,Etal "Title: Extractive Text Summarization using Deep Natural Language Fuzzy Processing " (April 2019)
4. Marc Everett Johnson "Title : Automatic Summarization Of Natural Language" (18 Dec 2018)
5. P. MAHALAKSHMI, N. SABIYATH FATIMA "TITLE :Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Technique"(February 18, 2022)
6. Tony Russell-Rose , Phil Gooch "Title : 2dSearch: a Visual Approach to Search Strategy Formulation"(August 2018)
7. Shpetim Sadriu1 "Title : Lecture Notes in Computer Science Technological Trends on Cognitive Virtual Assistants "(18 November 2020) Mohd Ibrahim Al-Qaoud, Rodina Ahmad "Title : Class diagram extraction from textual requirements using Natural language processing (NLP) techniques"(5 may 2015)
8. Kim Schouten, Flavius Frasincar "Title: Heracles: a Framework for Developing and Evaluating Text Mining Algorithms"(6 September 2020)
9. Florian Jungmann , G. Arnhold " Title : A Hybrid Reporting Platform for Extended RadLex Coding Combining Structured Reporting Templates and Natural Language Processing"(21 april 2020)