Medical Report Summarizer

Winta Yoseph, Kacie Do
Department of Information Science
DTSC 4050: Statistical Methods for Data Science and Analysis
Dr. Yang Zhang
May 7th, 2025

ABSTRACT

The healthcare industry is rapidly evolving, incorporating new technologies and innovation everyday. Electronic Health Records (EHRs) are a vital aspect of modern healthcare. Despite the technological advancements in healthcare, medical reports have a tendency of being too lengthy and difficult to interpret. The purpose of this project is to use Natural Language Processing (NLP) to develop a system that summarizes medical reports accurately and identifies key medical terms to improve efficiency in healthcare facilities. We used two models, ROUGE and Scipacy, to evaluate the generated medical summaries and identify key medical terminology. In addition, we obtained medical summaries and key terms from medical professionals for thorough comparison. After conducting an extensive evaluation, we received the following scores: ROUGE-1 F1: 0.5467, ROUGE-2 scores: 0.2761, and Longest Common Subsequence (LCS):0.2411. These scores indicate that the generated summaries moderately matched the sentence structures and captured a large amount of key words from the original summaries. Overall, these results support that our research provides an effective alternative for EHR usage, enhancing the efficiency of healthcare delivery and benefitting both healthcare providers and patients.

Keywords: Natural Language Processing, Text summarization, ROUGE Metric, TextRank, Scispacy

INTRODUCTION & MOTIVATION

Patient medical reports are often too long and complex which can delay medical diagnosis and treatment. The average physician spends approximately 16 minutes and 14 seconds reviewing Electronic Health Records (EHRs) for each patient. This does not take into account the time spent updating medical records (Lee, 2020).

Medical reports are subject to human errors such as grammar errors and typos. Because of this, healthcare professionals spend a lot of time trying to understand and confirm the contents of medical reports written by their colleagues. This can lead to misinterpretation which can compromise their ability to deliver quality healthcare. Depending on the health concern, the amount of time it takes to view and fully comprehend medical records can significantly impact patient outcomes (Aramaki et al., 2022). Our goal is to create a system that summarizes medical reports while prioritizing accuracy and readability in order to streamline healthcare processes.

RELATED WORK

There are a few common systems for Electronic Health Records (EHRs) that have incorporated Natural Language Processing (NLP) and generative Artificial Intelligence (AI) in their system. Many healthcare facilities utilize a company called Epic Systems to manage and store patient data. The system has a feature that analyzes information and events from earlier shifts and creates a concise summary for individuals on the next shift. It also embeds links for more detailed patient information within the summary (*Artificial Intelligence*, n.d.). This is a popular Electronic Health Record among medical professionals, as it adheres to the Health Insurance Portability and Accountability Act (HIPAA) regulations.

Another Electronic Health Record system is IBM Health Report Assistant, which is a database that allows healthcare professionals to upload and store medical reports. It transforms complex medical reports into summaries that can be easily interpreted by both patients and healthcare workers. Additionally, this system includes a chatbot feature that patients can use to ask questions regarding their health and/or treatment (*Infuse your product,* n.d.). This simplifies medical terminology for patients, so they can gain a better understanding of their diagnoses and treatments.

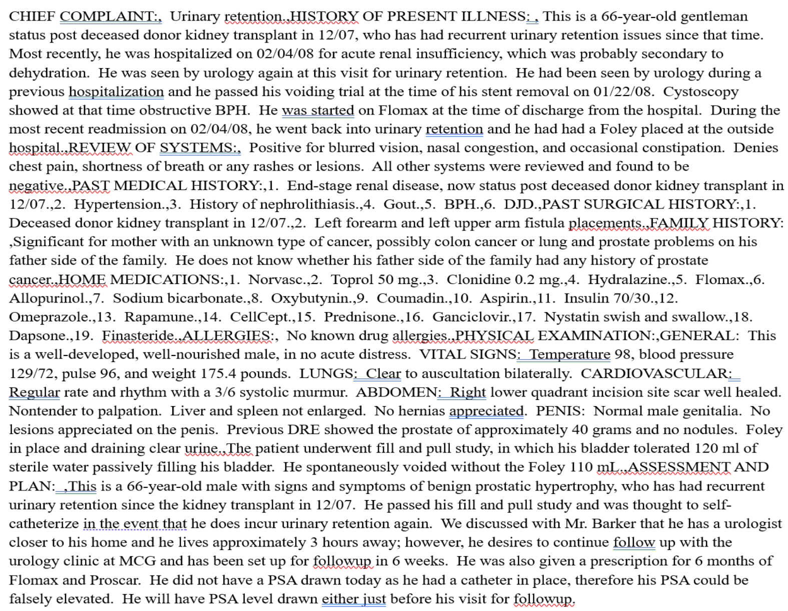
Amazon Comprehend Medical is another Electronic Health Record system that uses Natural Language Processing (NLP) to extract text from complex medical records. It creates simplified summaries of medical text while maintaining accuracy to expedite healthcare and medical insurance processes. It is also used to analyze population health data and identify inefficiencies within hospital settings (*Healthcare NLP,* n.d.).

Our proposed work leverages these existing systems to maximize efficiency and accuracy. We addressed the following problems:

(1) NLP does not work well with unstructured medical report data or free text.
(2) Some NLP's are limited to predetermined keywords.
(3) There is a lack of domain-specific awareness.

DATA

We extracted four medical reports out of a large dataset obtained from Kaggle.com and MTSamples.com.



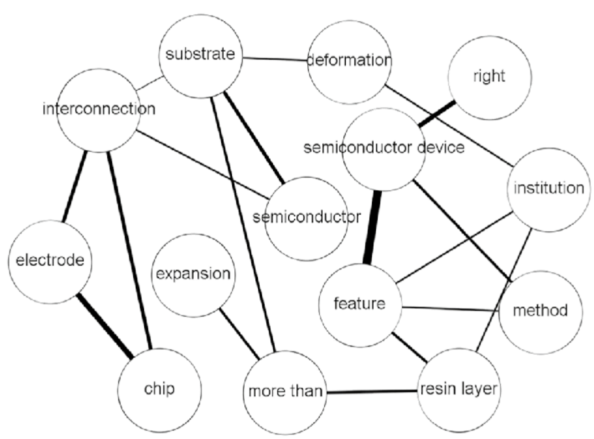
CHIEF COMPLAINT:, Urinary retention.,HISTORY OF PRESENT ILLNESS: , This is a 66-year-old gentleman status post deceased donor kidney transplant in 12/07, who has had recurrent urinary retention issues since that time. Most recently, he was hospitalized on 02/04/08 for acute renal insufficiency, which was probably secondary to dehydration. He was seen by urology again at this visit for urinary retention. He had been seen by urology during a previous hospitalization and he passed his voiding trial at the time of his stent removal on 01/22/08. Cystoscopy showed at that time obstructive BPH. He was started on Flomax at the time of discharge from the hospital. During the most recent readmission on 02/04/08, he went back into urinary retention and he had had a Foley placed at the outside hospital.,REVIEW OF SYSTEMS:, Positive for blurred vision, nasal congestion, and occasional constipation. Denies chest pain, shortness of breath or any rashes or lesions. All other systems were reviewed and found to be negative.,PAST MEDICAL HISTORY:,1. End-stage renal disease, now status post deceased donor kidney transplant in 12/07.,2. Hypertension.,3. History of nephrolithiasis.,4. Gout.,5. BPH.,6. DJD.,PAST SURGICAL HISTORY:,1. Deceased donor kidney transplant in 12/07.,2. Left forearm and left upper arm fistula placements.,FAMILY HISTORY: ,Significant for mother with an unknown type of cancer, possibly colon cancer or lung and prostate problems on his father side of the family. He does not know whether his father side of the family had any history of prostate cancer.,HOME MEDICATIONS:,1. Norvasc.,2. Toprol 50 mg.,3. Clonidine 0.2 mg.,4. Hydralazine.,5. Flomax.,6. Allopurinol.,7. Sodium bicarbonate.,8. Oxybutynin.,9. Coumadin.,10. Aspirin.,11. Insulin 70/30.,12. Omeprazole.,13. Rapamune.,14. CellCept.,15. Prednisone.,16. Ganciclovir.,17. Nystatin swish and swallow.,18. Dapsone.,19. Finasteride.,ALLERGIES:, No known drug allergies.,PHYSICAL EXAMINATION:,GENERAL: This is a well-developed, well-nourished male, in no acute distress. VITAL SIGNS: Temperature 98, blood pressure 129/72, pulse 96, and weight 175.4 pounds. LUNGS: Clear to auscultation bilaterally. CARDIOVASCULAR: Regular rate and rhythm with a 3/6 systolic murmur. ABDOMEN: Right lower quadrant incision site scar well healed. Nontender to palpation. Liver and spleen not enlarged. No hernias appreciated. PENIS: Normal male genitalia. No lesions appreciated on the penis. Previous DRE showed the prostate of approximately 40 grams and no nodules. Foley in place and draining clear urine.,The patient underwent fill and pull study, in which his bladder tolerated 120 ml of sterile water passively filling his bladder. He spontaneously voided without the Foley 110 mL.,ASSESSMENT AND PLAN: ,This is a 66-year-old male with signs and symptoms of benign prostatic hypertrophy, who has had recurrent urinary retention since the kidney transplant in 12/07. He passed his fill and pull study and was thought to self-catheterize in the event that he does incur urinary retention again. We discussed with Mr. Barker that he has a urologist closer to his home and he lives approximately 3 hours away; however, he desires to continue follow up with the urology clinic at MCG and has been set up for followup in 6 weeks. He was also given a prescription for 6 months of Flomax and Proscar. He did not have a PSA drawn today as he had a catheter in place, therefore his PSA could be falsely elevated. He will have PSA level drawn either just before his visit for followup.

Fig. 1. Example of a lengthy medical report containing complex language

METHODOLOGY

TextRank

TextRank is a graph based ranking algorithm that analyzes the relationships between words or phrases. It aims to create concise summaries without needing predefined keywords (Ramani et al., 2023).



[1]

Fig. 2. Depicts the implementation of classic TextRank model

The TextRank model helps to address problem 1 because it relies on word relationships

[1] A patent keywords extraction method using TextRank model with prior public knowledge - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/An-example-of-patent-TextRank-network_fig2_350472445 [accessed 29 Apr 2025]

rather than needing structured input. It also addressed problem 2 as there is no need for predefined vocabulary words. Additionally, it identifies the most important sentences based on context (Zaware et al., 2021).

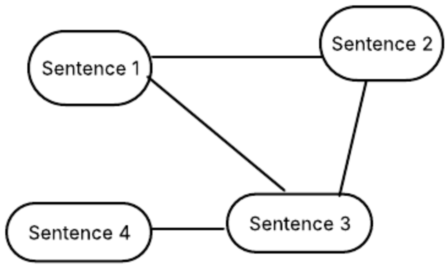
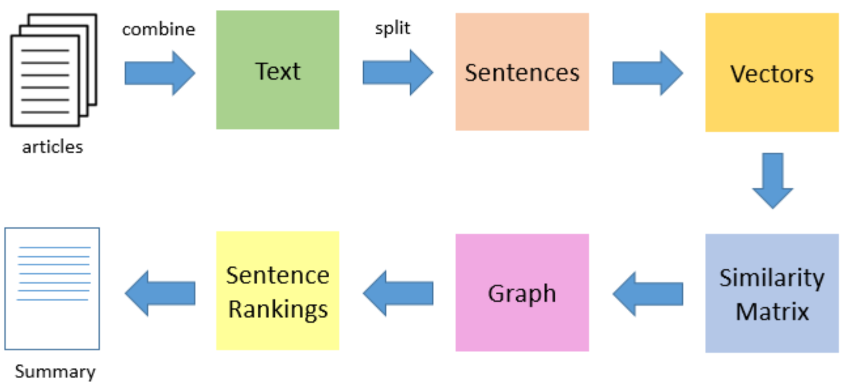


Fig. 3. Visualization of TextRank in our project



[2]

Fig. 4. Workflow of TextRank

Scispacy

After summarizing the medical text using TextRank, we implemented Scispacy, an extension built on top of Spacy, which is trained on biomedical and scientific texts. This helps address problem 3 by adding domain specific awareness to the summaries. It detects and extracts named entities such as diseases and chemicals, recognizing medical terminology using NER tags (Agarwal et al., 2024).

[2] https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

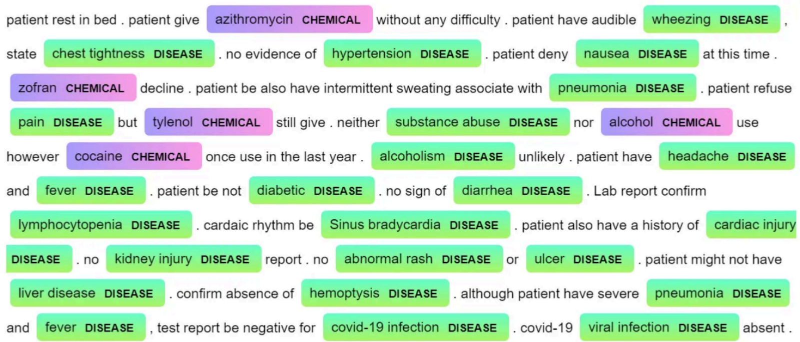
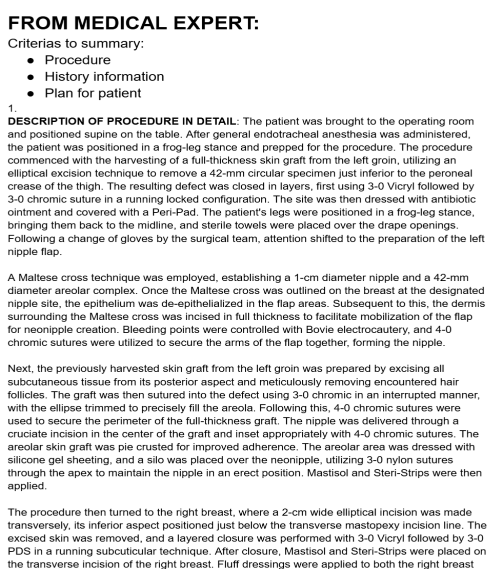


Fig. 5. Example of Sciscpacy[3]

EXPERIMENTS & RESULTS

During our research process, we reached out to medical professionals to gather their expert opinion on the quality of the original medical reports from our dataset. Then, we asked them to summarize four medical reports in their own words. We used their responses as the "gold standard" for the ROUGE model. In our evaluation, we compared the generated summary to the summary provided by medical professionals. This allowed us to measure the accuracy of the generated summaries. In addition to the summarization, we asked the medical professionals to identify and highlight any key medical terms such as health conditions and medications. This information allowed us to evaluate the Scispacy model.



FROM MEDICAL EXPERT:

Criterias to summary:

- Procedure
- History information
- Plan for patient

1.

DESCRIPTION OF PROCEDURE IN DETAIL: The patient was brought to the operating room and positioned supine on the table. After general endotracheal anesthesia was administered, the patient was positioned in a frog-leg stance and prepped for the procedure. The procedure commenced with the harvesting of a full-thickness skin graft from the left groin, utilizing an elliptical excision technique to remove a 42-mm circular specimen just inferior to the peroneal crease of the thigh. The resulting defect was closed in layers, first using 3-0 Vicryl followed by 3-0 chromic suture in a running locked configuration. The site was then dressed with antibiotic ointment and covered with a Peri-Pad. The patient's legs were positioned in a frog-leg stance, bringing them back to the midline, and sterile towels were placed over the drape openings. Following a change of gloves by the surgical team, attention shifted to the preparation of the left nipple flap.

A Maltese cross technique was employed, establishing a 1-cm diameter nipple and a 42-mm diameter areolar complex. Once the Maltese cross was outlined on the breast at the designated nipple site, the epithelium was de-epithelialized in the flap areas. Subsequent to this, the dermis surrounding the Maltese cross was incised in full thickness to facilitate mobilization of the flap for neonipple creation. Bleeding points were controlled with Bovie electrocautery, and 4-0 chromic sutures were utilized to secure the arms of the flap together, forming the nipple.

Next, the previously harvested skin graft from the left groin was prepared by excising all subcutaneous tissue from its posterior aspect and meticulously removing encountered hair follicles. The graft was then sutured into the defect using 3-0 chromic in an interrupted manner, with the ellipse trimmed to precisely fill the areola. Following this, 4-0 chromic sutures were used to secure the perimeter of the full-thickness graft. The nipple was delivered through a cruciate incision in the center of the graft and inset appropriately with 4-0 chromic sutures. The areolar skin graft was pie crusted for improved adherence. The areolar area was dressed with silicone gel sheeting, and a silo was placed over the neonipple, utilizing 3-0 nylon sutures through the apex to maintain the nipple in an erect position. Mastisol and Steri-Strips were then applied.

The procedure then turned to the right breast, where a 2-cm wide elliptical incision was made transversely, its inferior aspect positioned just below the transverse mastopexy incision line. The excised skin was removed, and a layered closure was performed with 3-0 Vicryl followed by 3-0 PDS in a running subcuticular technique. After closure, Mastisol and Steri-Strips were placed on the transverse incision of the right breast. Fluff dressings were applied to both the right breast

Fig. 6. One medical summary provided by a medical professional

[3] https://medium.com/@MansiKukreja/clinical-text-negation-handling-using-negspacy-and-scispacy-233ce69ab2ac

RESULTS

In our research, we utilized the long medical transcriptions. We applied different python language packages for preprocessing the text. Then, we created two machine learning models, TextRank and Scispacy, to construct the comprehensive process. To evaluate the performance of these models, the ROUGE score and evaluation metric were used for TextRank and Scispacy respectively.

The result after applying models in the lengthy medical report in Fig. 1:

Phase 1: Apply model TextRank

Output

```
 ' End-stage renal disease, now status post deceased donor kidney transplant in 12/07.,2. He spontaneously voided without the Foley 110 mL.,ASSESSMENT AND PLAN:  ,This is
a 66-year-old male with signs and symptoms of benign prostatic hypertrophy, who has had recurrent urinary retention since the kidney transplant in 12/07. Insulin
70/30.,12. He had been seen by urology during a previous hospitalization and he passed his voiding trial at the time of his stent removal on 01/22/08. No lesions
appreciated on the penis. He will have PSA level drawn either just before his visit for followup. During the most recent readmission on 02/04/08, he went back into
urinary retention and he had had a Foley placed at the outside hospital.,REVIEW OF SYSTEMS:,  Positive for blurred vision, nasal congestion, and occasional constipation.
Flomax.,6. Finasteride.,ALLERGIES:,  No known drug allergies.,PHYSICAL EXAMINATION:,GENERAL:  This is a well-developed, well-nourished male, in no acute distress. We
discussed with Mr. Barker that he has a urologist closer to his home and he lives approximately 3 hours away; however, he desires to continue follow up with the urology
clinic at MCG and has been set up for followup in 6 weeks. LUNGS:  Clear to auscultation bilaterally. BPH.,6. DJD.,PAST SURGICAL HISTORY:,1. Left forearm and left upper
arm fistula placements.,FAMILY HISTORY:  ,Significant for mother with an unknown type of cancer, possibly colon cancer or lung and prostate problems on his father side of
the family. History of nephrolithiasis.,4.']
```

Fig. 7. Shows an output summary generated by the model

Model performance

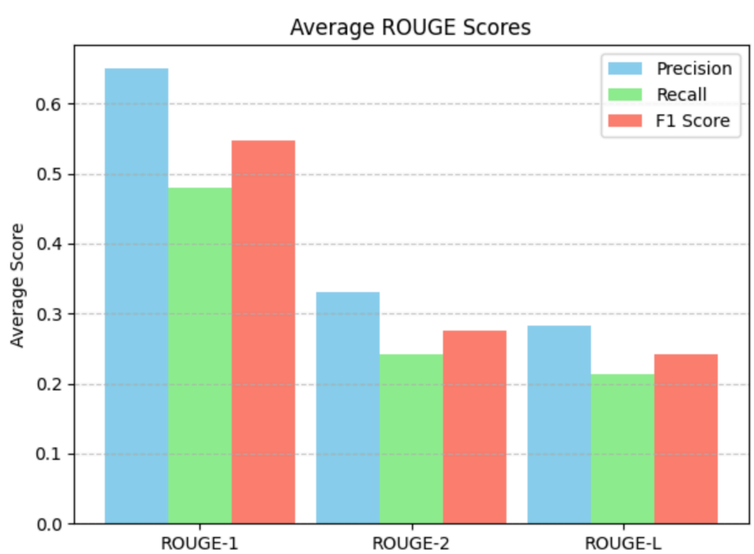


Fig. 8. Shows the Average ROUGE Scores of the TextRank

Phase 2: Apply model Scispacy

Output

End-stage renal disease **DISEASE** , now status post deceased donor kidney transplant in 12/07.,2. He spontaneously voided without the Foley 110 mL.,ASSESSMENT AND PLAN: ,This is a 66-year-old male with signs and symptoms of benign prostatic hypertrophy **DISEASE** , who has had recurrent urinary retention since the kidney transplant in 12/07. Insulin 70/30.,12. He had been seen by urology during a previous hospitalization and he passed his voiding trial at the time of his stent removal on 01/22/08. No lesions appreciated on the penis. He will have PSA level drawn either just before his visit for followup. During the most recent readmission on 02/04/08, he went back into urinary retention and he had had a Foley placed at the outside hospital.,REVIEW OF SYSTEMS:, Positive for blurred vision **DISEASE** , nasal congestion **DISEASE** , and occasional constipation **DISEASE** . Flomax.,6. Finasteride.,ALLERGIES:, No known drug allergies.,PHYSICAL EXAMINATION:,GENERAL: This is a well-developed, well-nourished male, in no acute distress. We discussed with Mr. Barker that he has a urologist closer to his home and he lives approximately 3 hours away; however, he desires to continue follow up with the urology clinic at MCG and has been set up for followup in 6 weeks. LUNGS: Clear to auscultation bilaterally. BPH.,6. DJD.,PAST **DISEASE** SURGICAL HISTORY:,1. Left forearm and left upper arm fistula **DISEASE** placements.,FAMILY HISTORY: ,Significant for mother with an unknown type of cancer **DISEASE** , possibly colon cancer **DISEASE** or lung and prostate **DISEASE** problems on his father side of the family. History of nephrolithiasis.,4.

Fig. 9. Shows an output of the Scispacy model

Model performance

```
Precision: 0.7
Recall: 0.21212121212121213
F1 Score: 0.3255813953488372
```

Fig. 10. Shows the performance result of the Scispacy model

DISCUSSION & FINDINGS

Model TextRank

The F1 Score balances both the precision and recall scores using the harmonic mean. The average ROUGE-1 F1 is 0.5467, indicating that the system captured a large amount of keywords from the original summaries. ROUGE-2 scores (0.2761) are typically lower than ROUGE-1 as bigrams are harder to match, indicating a moderate level of matching phrases between the generated and original summaries. Longest Common Subsequence (LCS) is 0.2411, indicating that the generated summaries moderately matched the sentence structure of the original summaries (*Rouge Evaluation Metric* 2025).

TextRank helps reduce noise and focuses on key sentences, but it also removes some essential biomedical entities that the summarization algorithm doesn't consider central.

Model Scispacy

The precision score is pretty high, showing that most of the identified entities in the summary are correct. However, the recall score is low, showing a loss in term coverage, which is especially critical in medical contexts where completeness matters.

LIMITATIONS & FUTURE WORK

TextRank is extractive purely. This leads to the fact that it might generate summaries that lack coherence. Additionally, typo errors and noisy data still present in the output, post-processing step is essential for future development. Most importantly, due to the limited availability of medical experts, we were only able to ask our experts for a summary on a few samples from the dataset. Therefore, this project focuses on developing and demonstrating the workflow of the entire process, rather than statistically proving its effectiveness. To improve accuracy, collaboration with more medical experts to generate high quality reference summaries is necessary for the future work.

CONCLUSION

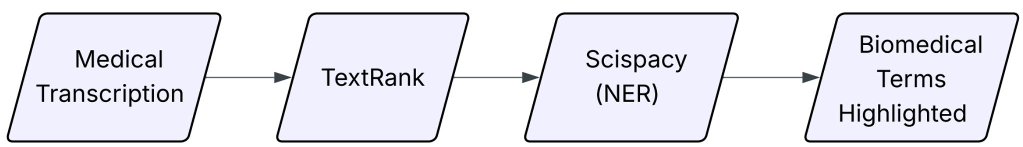


Fig. 11. The workflow of the system

The approach in this project offers a transparent pipeline, making it useful and easier for medical professionals to understand why certain contents are prioritized in summary. It also helps accelerate workflows and assist in medical decision making by providing more reliable information. Our approach would be ideal for fast or high-precision extraction (e.g. quick reviews), but not ideal where full coverage is required (e.g. in comprehensive patient history analysis).

REFERENCES

Agarwal, T., Singh, B., & Chugh, A. (2024). *MedBot: A novel sequential pipeline for context recognition based on SciSpacy and Med7 entity recognizer*. In 2024 Second International Conference on Advanced Computing & Communication Technologies (ICACCTech) (pp. 875–879). IEEE. https://doi.org/10.1109/ICACCTech65084.2024.00143

Aramaki, E., Wakamiya, S., Yada, S., & Nakamura, Y. (2022, June). *Natural language processing: From bedside to everywhere*. Yearbook of medical informatics. https://pmc.ncbi.nlm.nih.gov/articles/PMC9719781/

Artificial Intelligence. Epic. (n.d.). https://www.epic.com/software/ai/

Healthcare NLP - extract health data - amazon comprehend medical - AWS. (n.d.). https://aws.amazon.com/comprehend/medical/

Infuse your product with artificial intelligence from IBM. IBM Digital Self-serve co-create experience. (n.d.). https://dsce.ibm.com/wizard/watsonx/results/watsonx-health-report-assistant

Lee, B. Y. (2020, January 14). *How doctors may be spending more time with electronic health records than patients*. Forbes. https://www.forbes.com/sites/brucelee/2020/01/13/electronic-health-records-here-is-how-much-time-doctors-are-spending-with-them/

Rouge Evaluation Metric. (2025, March 7). https://www.ibm.com/docs/en/watsonx/saas?topic=metrics-rouge

Ramani, K., Bhavana, K., Akshaya, A., Harshita, K. S., Thoran Kumar, C. R., & Srikanth, M. (2023). *An explorative study on extractive text summarization through K-means, LSA, and TextRank*. In 2023 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET) (pp. 1–6). IEEE. https://doi.org/10.1109/WiSPNET57748.2023.10134303

Zaware, S., Patadiya, D., Gaikwad, A., Gulhane, S., & Thakare, A. (2021). *Text summarization* *using TF-IDF and TextRank algorithms*. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1399–1407). IEEE. https://doi.org/10.1109/ICOEI51242.2021.9453071

APPENDIX

Code Implementation

For readers to refer and experiment with the project's source code, we have deployed the code in the Google Colab Notebook. You can access and run the code directly without installing any additional software. The link below will lead you to the notebook

https://colab.research.google.com/drive/19r1sGeRTn2yj3y8_FgOfSUH7nDGrF4Ke?usp=sharing