Project 33

Mining the University of Oulu Hospital ESKO Database

Mohammad Fahad Khalid (2307232) & Khuzaima Arham (2310591)

Project on Github:
https://github.com/fahadchauhan/Mining-University-of-Oulu-Hospital-ESKO-database

Abstract---This study delves into the University of Oulu Hospital's ESKO database, utilizing Natural Language Processing (NLP) techniques to grasp clinician updates and patient records. It extensively examines medical ontologies, conducts narrative analysis, and visualizes data to gain valuable insights into the database's medical terms. The primary tool for terminology analysis is the Unified Medical Language System (UMLS). The goal is to improve the understanding of medical language in healthcare records, ultimately enhancing patient care and data management in healthcare settings.

---+---

Keywords— Healthcare Records, Python, NLTK, NLP, Medical Vocabulary Analysis, Medical Vocabulary Analysis, Clinician Recommendations

I. *Introduction*

The advancement in electronic health records (EHR) and the utilization of healthcare data have become instrumental in healthcare management, diagnosis, and treatment plans[1]. The growing adoption of electronic systems has led to vast amounts of patient data generated, creating a new challenge in mining and analyzing such extensive information. This phenomenon has driven the need for innovative approaches and techniques to extract meaningful insights, particularly concerning the medical terminologies and narratives used within the healthcare domain.

The University of Oulu Hospital's ESKO database provides a rich repository of patient records, offering a significant opportunity to explore these narratives and terminology within the medical domain. The analysis of this data aims to contribute to the comprehension of medical vocabularies used by healthcare professionals, providing critical insights into the language employed in patient records and clinician interactions.

This paper conducts a concise review of the existing literature surrounding the application of Natural Language Processing (NLP) techniques in analyzing medical text and uncovering insights from healthcare data[2]. Previous studies have highlighted the potential challenges in data management, specifically in processing large volumes of medical narratives, extracting pertinent medical terminologies, and identifying critical nuances in healthcare records. By leveraging the Unified Medical Language System (UMLS)[3], previous research has shown promise in extracting and understanding medical terms within patient records[4]. However, there remains a gap in understanding the specific usage and adoption of medical terms by healthcare professionals, particularly in a clinical setting.

The aim of this project is to bridge this gap by conducting an in-depth analysis of medical terminologies in the ESKO database, specifically examining the narratives and terms used by clinicians. This exploration seeks to improve medical data management and enhance patient care by understanding and extracting critical medical terminologies, thereby providing a better comprehension of the nuances in clinical narratives. Through this endeavour, the study endeavours to shed light on the challenges, constraints, and potential solutions in mining and understanding healthcare data, ultimately contributing to the optimization of healthcare information systems and improving patient care outcomes.

II. *Methodology*

A. Theoretical Foundation

There are two of the main theoretical foundations in the project. Following are the foundations of the methodology used in the project.

a. Introduction to NLP in Healthcare Analysis:

Natural Language Processing (NLP) forms the cornerstone of interpreting, analyzing, and categorizing extensive unstructured medical text data. Through techniques such as term categorization, translation, statistical analysis, vocabulary extraction, correlation, and in-depth analysis, NLP plays a pivotal role in transforming unstructured data into structured formats. These methods are instrumental in unveiling trends, patterns, and vital medical terminologies, thereby significantly enhancing patient care and enabling more informed decision-making within healthcare systems.

b. Unified Medical Language System (UMLS) Exploration:

The Unified Medical Language System (UMLS) serves as a fundamental resource in the healthcare domain, offering a cohesive platform that integrates a diverse range of biomedical terminologies and standards. By consolidating various biomedical vocabularies, classifications, and coding systems, the UMLS plays a pivotal role in fostering a comprehensive and consistent understanding of medical terminologies. It acts as a valuable tool for the extraction and comprehension of medical concepts and terms, providing an extensive inventory and relationships among different terminologies. This system aids in harmonising diverse terminologies, facilitating interoperability and ensuring a shared understanding of healthcare terminologies across different information systems and applications.

B. NLP Pipeline

The adopted multi-tiered methodology facilitated a comprehensive analysis of doctor recommendations. This approach aimed to illuminate the language utilized, the prevalence of medical vocabulary, and the underlying themes or sentiments encapsulated within the advice offered to patients. The process involved various stages of data processing, analysis, and visualization to derive meaningful insights from the narrative texts of medical recommendations.

a. Data Preprocessing:

The data preprocessing phase primarily focused on filtering and refining the available narrative texts by excluding records with missing or empty narrative fields. This meticulous approach ensured that only complete narrative texts were considered for further analysis. The process concentrated on optimizing the narrative texts for accuracy and relevance, utilizing text normalization techniques to eliminate new lines, extra spaces, punctuations, stop words, and numbers. These steps aimed to enhance the coherence and usability of the narrative content for subsequent Natural Language Processing (NLP) analysis while maintaining the integrity of the original dataset, ensuring a comprehensive yet targeted approach to data preparation.

b. Text Translation

The translation process involved using the Google Translation API to convert Finnish (FI) narrative texts to English (EN). To comply with the API's restrictions on the number of queries per second, a deliberate 0.2-second delay was introduced between translation requests. This strategic pause effectively regulated the rate of API calls, ensuring precise and consistent translations of language-specific content while adhering to the prescribed usage parameters. By managing translation calls with this deliberate timing, the team successfully circumvented the API limitations, ensuring the accuracy and consistency of translated narrative text for further analysis.

c. Statistical Analysis

For the statistical analysis, the team computed various narrative text statistics based on the translated content obtained from the Google API translation. Metrics such as the mean word count and standard deviation, along with skewness and kurtosis, were calculated to gain insights into the distribution, shape, and characteristics of the translated narratives. These statistical measurements were crucial in understanding the average length of the translated texts, as well as the extent of variability, asymmetry, and tails in the distribution of word counts. This comprehensive statistical analysis provided valuable insights into the nature and structure of the narrative content, facilitating a deeper understanding of the translated data.

d. WordCloud Generation

For the creation of WordCloud representations, the team utilized the WordCloud library to visually depict the frequency of words within the translated narratives obtained through the Google Translation API. This involved generating a visual representation where each word's size corresponded to its frequency within the narratives. The WordCloud representations offered an intuitive and comprehensive view of the most frequently occurring words, enabling easy interpretation and identification of dominant or recurring terms within the translated narrative texts. This visual method facilitated the identification of prevalent terms and emphasized the distribution of word frequency, providing a clear insight into the most common linguistic elements present in the translated narratives.

e. Empath Categorization

The utilization of the Empath tool played a critical role in categorizing the translated narrative content, providing a comprehensive comprehension of prevalent themes and the emotional context embedded within the text. Post-categorization, is a filtering procedure on categories of the entire text, aiming to retain the most significant and prevalent thematic categories for a focused and insightful analysis.

f. Medical Terminology Extraction

The process of extracting medical terminologies from narrative texts commenced with the integration of the Unified Medical Language System (UMLS), an extensive repository housing a wide spectrum of medical concepts and terminologies. Initially, the team attempted to perform this extraction using the UMLS API[5], striving to match semantic types with the relevant tree numbers. However, due to time constraints and efficiency concerns, an alternative approach was adopted. The team opted for downloading the UMLS data and establishing a local database. This decision aimed to expedite the querying process and facilitate a quicker and more focused retrieval of medical terminologies from the narrative texts.

Subsequently, the string-matching process aligned the contents of the narrative texts with the concepts stored in the UMLS, aiming to correlate terms present in the narratives with the corresponding concepts within the UMLS database. After establishing these connections, the focus turned to discerning the semantic types associated with each concept. This categorization was primarily accomplished by retrieving Concept Unique Identifiers (CUIs), which laid the foundation for filtering out non-medical terms.

For a more refined filtering mechanism, the team manually curated and condensed the list of semantic types. This curation process was based on the Term Unique Identifiers (TUIs) within the UMLS. By selecting and shortlisting TUIs specific to medical terminologies, the process aimed to improve the accuracy and relevance of the extracted medical vocabulary within the narrative texts.

The objective of these meticulous methodologies was to ensure the precision and relevance of the identified medical terminologies. By focusing solely on terms directly associated with the medical field, the approach aimed to improve the accuracy and applicability of the extracted medical vocabulary. In a bid to offer transparency and facilitate further research, the resulting list of tree numbers and TUIs utilized in the curation process will be provided alongside the project documentation. This resource will serve as a valuable guide for researchers and interested parties, aiding in understanding the methodology applied in the extraction of medical terminologies from the narrative texts.

g. Correlation Analysis: Medical Vocabulary and Word Count in Narrative Texts

In the correlation analysis of medical vocabulary and word count in the narrative texts, we sought to investigate any potential correlation between the usage of medical vocabulary and the length of the narrative texts. This analysis aimed to determine whether there was a relationship between the frequency of medical terms and the number of words in each narrative. To represent this, a histogram was generated, depicting the total number of narrative texts along with their average size and standard deviation concerning the most commonly used medical vocabulary.

h. Comprehensive Analysis of Doctor Recommendations

This comprehensive analysis delineates the multi-step approach employed to investigate and comprehend the language, medical vocabulary, and underlying themes in doctor recommendations present within the narrative texts.

i. Doctor Recommendation Identification:

In the process of identifying doctor recommendations within the narrative texts, a verb dictionary was constructed by leveraging Part-of-Speech (POS) tags, specifically focusing on verbs categorized as "VB," "VBP," and "VBZ." Sentences containing these verbs were extracted to pinpoint various keywords and phrases indicative of recommendations. The goal was to capture diverse forms of advice related to tests, lifestyle adjustments, revisits to the clinic, referrals to services, or specific treatments. The extraction of this information was facilitated by employing the Spacy model "en_core_web_sm.

ii.WordCloud for Recommendation Word Usage:

To understand the emphasis of certain terms within the doctor recommendations, We utilized a WordCloud, generated to visually represent the frequency of words in these recommendations. This approach aims to capture the most frequent and reiterated concepts, enabling a clearer understanding of the prominent terms in the recommendations.

iii. UMLS for Medical Vocabulary Proportion:

Utilizing the Unified Medical Language System (UMLS), terms extracted from the recommendations were mapped to the UMLS database to measure the prevalence of medical terminologies. This process aimed to quantify the usage of medical vocabulary within the doctor's advice, providing insights into the medical language incorporated within the recommendations.

iii. Empath Categorization of Recommendation Context:

The Empath categorization process was repeated to comprehend the primary themes, sentiments, and emotional context conveyed within the doctor's recommendations. Analyzing these statements provided categorized insights into the prevalent themes, sentiments, and the underlying emotional context within the recommendations.

i. Graphical User Interface (GUI):

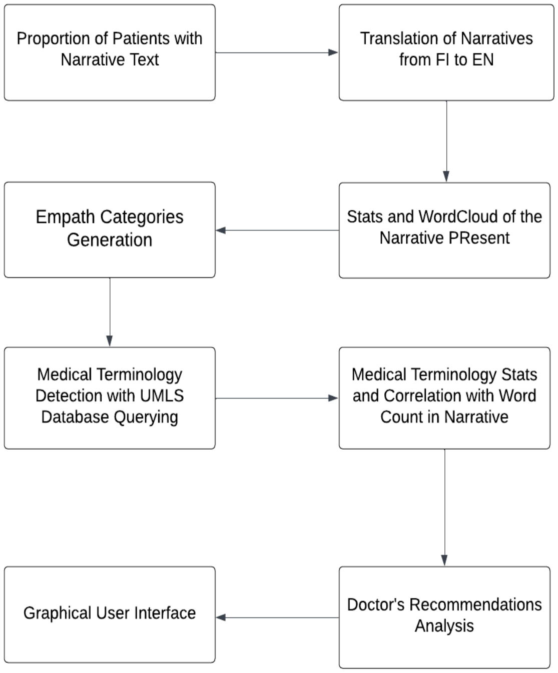
Developed an intuitive interface to demonstrate the execution of each of the previously mentioned tasks. This graphical interface aims to simplify the assessment process for the evaluator or external end-users.

C. Implementation Details

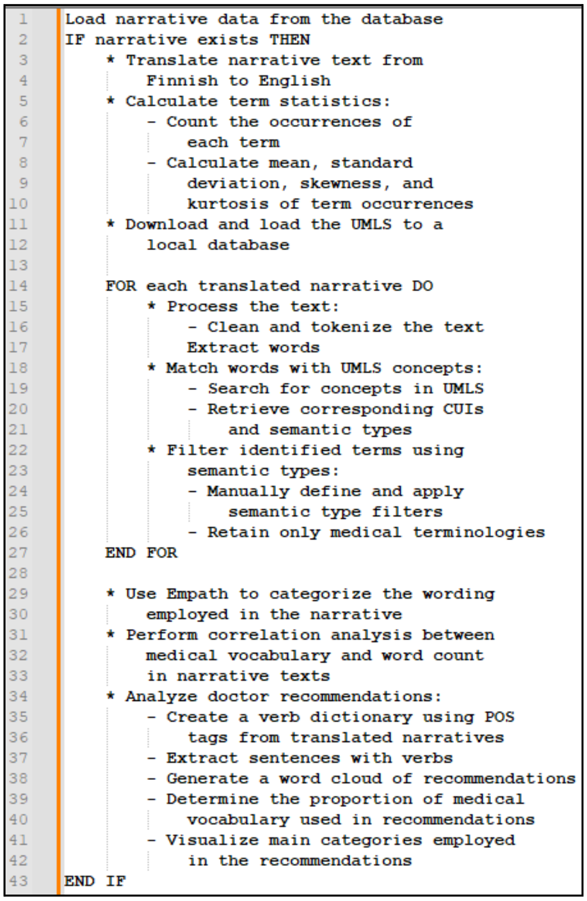
a. Coding Approach:

The methodology for this project predominantly relied on Python and various libraries that offered specialized functionalities. The core libraries involved Pandas, primarily employed for data manipulation, cleaning, and structuring the datasets. NLTK (Natural Language Toolkit) played a pivotal role in facilitating multiple text processing tasks, especially in aspects related to text tokenization, part-of-speech tagging, and syntactic analysis. The Empath library was used for empathetic text analysis, providing insights into the emotional and thematic context within the narrative texts. The googletrans library was utilized to connect with the Google Translation API, enabling the translation of Finnish text to English, broadening the comprehensibility of the data. Additionally, scipy aided in statistical computations, spacy contributed to advanced natural language processing techniques, and WordCloud supported the creation of visually representative word frequencies. The Python programming language was central to orchestrating these libraries, enabling a comprehensive approach to analyzing and processing the narrative texts.

b. Block Diagram:


Proportion of Patients with Narrative Text
Translation of Narratives from FI to EN
Stats and WordCloud of the Narrative PResent
Empath Categories Generation
Medical Terminology Detection with UMLS Database Querying
Medical Terminology Stats and Correlation with Word Count in Narrative
Doctor's Recommendations Analysis
Graphical User Interface

c. Pseudo-Code:



```
Load narrative data from the database
IF narrative exists THEN
    * Translate narrative text from
      Finnish to English
    * Calculate term statistics:
        - Count the occurrences of
          each term
        - Calculate mean, standard
          deviation, skewness, and
          kurtosis of term occurrences
    * Download and load the UMLS to a
      local database

    FOR each translated narrative DO
        * Process the text:
            - Clean and tokenize the text
            Extract words
        * Match words with UMLS concepts:
            - Search for concepts in UMLS
            - Retrieve corresponding CUIs
              and semantic types
        * Filter identified terms using
            semantic types:
            - Manually define and apply
              semantic type filters
            - Retain only medical terminologies
    END FOR

    * Use Empath to categorize the wording
      employed in the narrative
    * Perform correlation analysis between
      medical vocabulary and word count
      in narrative texts
    * Analyze doctor recommendations:
        - Create a verb dictionary using POS
          tags from translated narratives
        - Extract sentences with verbs
        - Generate a word cloud of recommendations
        - Determine the proportion of medical
          vocabulary used in recommendations
        - Visualize main categories employed
          in the recommendations
END IF
```

III. Results and Discussions

A. Data Preprocessing:

Aimed to visualize the presence of narrative texts in patient records. The analysis revealed that out of a total of 1,293,359 records, only 24,352 entries (approximately 1.88%) included narrative text, while the vast majority—specifically 1,269,007 records (around 98.11%)—were devoid of any narrative text. This significant contrast was visually represented in Figure 1, titled 'Proportion of Patients with Narrative Text'.

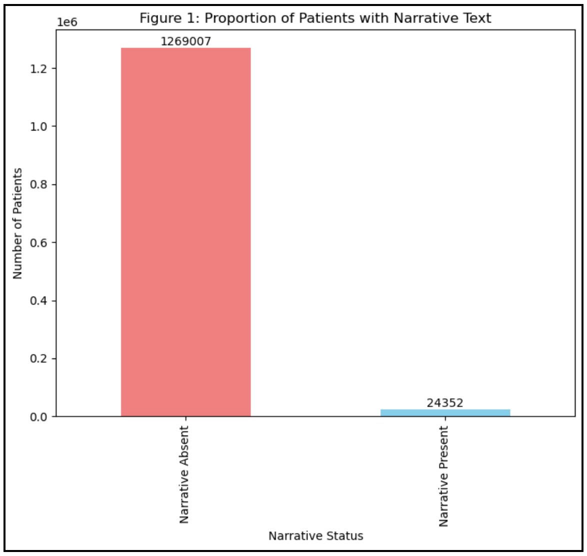


Figure 1: Proportion of Patients with Narrative Text

B. Statistical Analysis

The statistical analysis performed on the translated content from the Google API translation provided substantial insights into the nature and structure of the narrative texts. The mean word count calculated was approximately 63.34 words, with a significant standard deviation of 79.69 (Figure 2). These findings indicate a notable variability in the length of the translated texts. Moreover, the calculated skewness and kurtosis, reported as 3.19 and 17.28, respectively (Figure 3), indicate a considerable asymmetry and the presence of outliers in the distribution of word counts.

These statistical measurements not only revealed the average length of the translated texts but also highlighted the variation, asymmetry, and distribution characteristics of the word count within the narratives. The high standard deviation and significant values for skewness and kurtosis signify the diverse nature and spread of the word count data, suggesting a varied distribution of word lengths across the translated narratives.

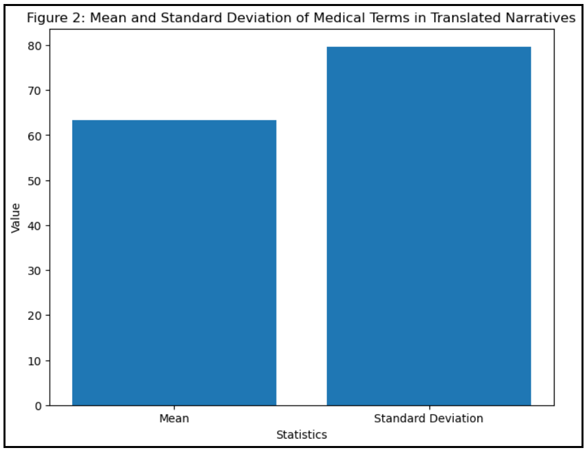


Figure 2: Mean and Standard Deviation of Medical Terms in Translated Narratives

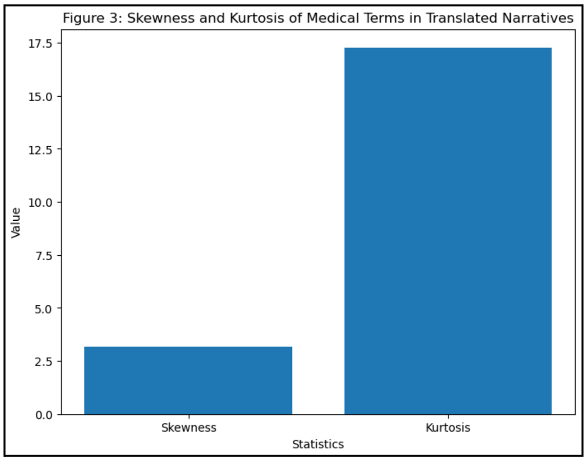


Figure 3: Skewness and Kurtosis of Medical Terms in Translated Narratives

C. WordCloud Generation

The WordCloud representation generated from the translated narratives offered essential insights into the most frequent terms within the dataset. From WordCloud, several prominent terms were identified, including "treatment," "patient," "found," "surgery," "situation," "control," and "health centre." These frequently occurring terms, visualized in the WordCloud (Figure 4), highlighted the prevalent topics and recurring themes present in the translated narratives.

The emphasis on specific terms in the WordCloud is indicative of the most commonly used vocabulary within the translated texts. Terms such as "treatment" and "patient" indicate a focus on medical interventions and patient-related discussions. Moreover, words like "surgery," "health-centre," and "control" suggest aspects related to medical procedures, healthcare facilities, and disease management.

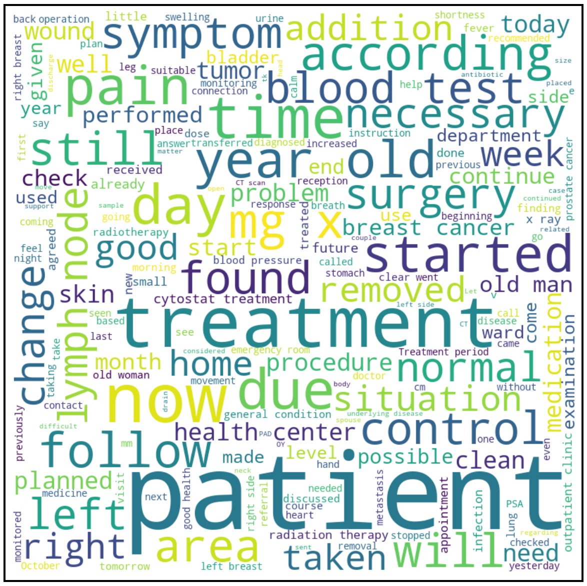


Figure 4: WordCloud representation of the translated narratives

The WordCloud visualization facilitated an immediate understanding of the dominant topics within the narratives, offering a condensed overview of the most prevalent terms. It proved to be an effective tool for identifying and interpreting the most frequently used words, contributing to a better understanding of the thematic content within the translated narratives.

D. Empath Categorization

The Empath categorization offered a comprehensive understanding of the prevailing themes and the emotional context embedded within the translated narratives. After employing a post-categorization filtering process that targeted categories with proportions exceeding 0.001 of the entire text, we identified a maximum value of category proportions at 0.0399. The filtering threshold of 0.001 was selected to focus on the most substantial thematic categories for a more detailed and insightful analysis.

Subsequently, to streamline the categories further, the threshold was adjusted to 0.003, resulting in a more focused categorization. Figure 5 illustrated the Proportion of Empath Categories Exceeding 0.003 in Narratives. This filtering presented several notable categories, including "health," "medical_emergency," "healing," "help," "injury," "body," "liquid," "school," "pain," "communication," "meeting," "negative_emotions," and "hearing."

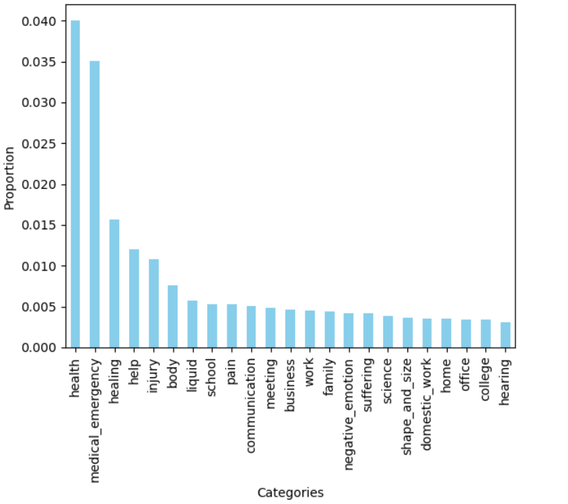


Figure 5: Proportion of Empath Categories Exceeding 0.003 in Narratives

These top categories represented a range of themes encompassing health-related issues, emotional concerns, and communication dynamics. The prevalence of categories like "health," "medical_emergency," and "pain" indicates the dominant focus on medical concerns and emergencies. Furthermore, terms such as "communication," "meeting," and "negative_emotions" suggest elements of interpersonal interactions and emotional states prevalent in the narratives.

The adjusted filtering threshold refined the categories, offering a clearer and more precise view of the most prevalent thematic elements within the narratives, which proved instrumental in understanding the most significant themes and emotional context embedded within the translated texts.

E. Medical Terminology Extraction

The statistical examination performed on the identified medical terms within the narratives revealed compelling findings. The mean number of medical terms across these narratives averaged at 17.03, showcasing considerable variability indicated by the standard deviation of approximately 20.73.

Additionally, insights derived from the skewness and kurtosis measurements shed light on the distribution characteristics of medical terms in these narratives. A skewness value of 2.93 suggests a notable asymmetry in the term distribution, potentially indicating a more pronounced occurrence of certain medical terminologies. The kurtosis value, at 13.92, implies a higher presence of outliers or infrequent terminologies, contributing to a distribution more concentrated around the mean.

Figure 6: Mean and Standard Deviation of Medical Terms of Translated Narratived

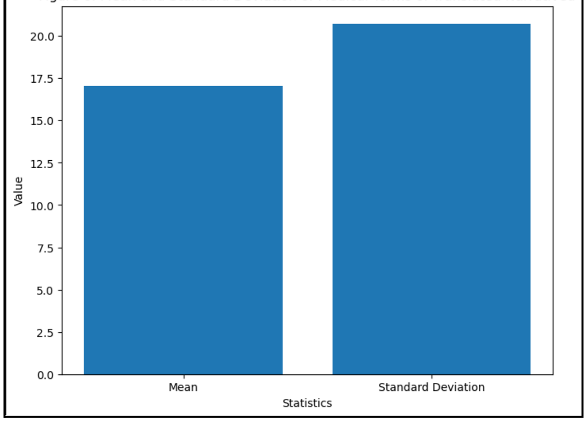
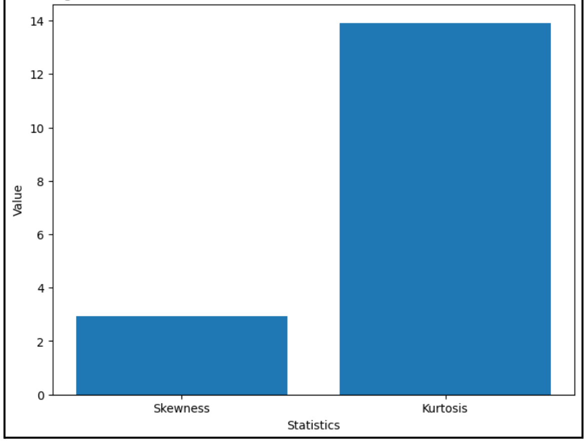


Figure 7: Skewness and Kurtosis of Medical Terms of Translated Narratived



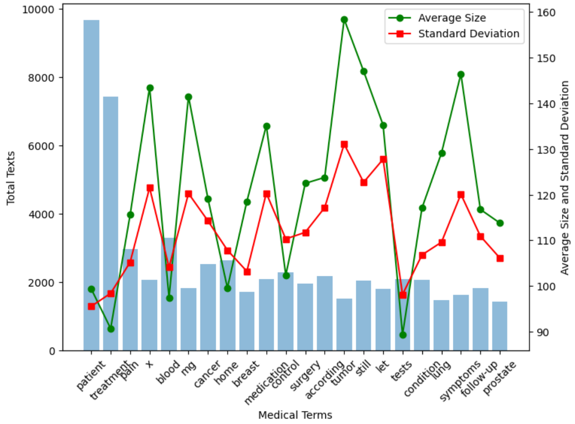
The substantial variance observed in the prevalence of medical terms indicates a diverse spectrum of terminologies and their variable usage within the analyzed texts. These findings provide valuable insights into the abundance and distribution of medical language in the narrative content, contributing to a deeper comprehension of healthcare-related contexts covered in the narratives.

These statistical insights, including the mean, standard deviation (Figure 6), skewness, and kurtosis (Figure 7) measurements, offer a detailed understanding of the medical term distribution and their prevalence in the narrative texts.

F. Correlation Analysis: Medical Vocabulary and Word Count in Narrative Texts

The correlation analysis aimed to explore the relationship between the occurrence of medical vocabulary and the length of the narrative texts. By investigating the frequency of medical terms within each narrative, we sought to discern whether there was a correlation between the usage of medical vocabulary and the word count in each narrative. The analysis was visualized using a histogram, showcasing the total number of narrative texts and their average size, along with the standard deviation concerning the most frequently used medical vocabulary. The graphical representation in Figure 8 illustrates this analysis.

Figure 8: Medical Vocabulary vs. Text Statistics (Medical term Frequency > 2000)



Upon analyzing the graph, various medical vocabulary terms stood out, each showing a range of occurrences and variations in their usage within the narratives. The prevalent medical terms such as "patient," "treatment," "x," "blood," "cancer," and others appeared to exhibit a substantial frequency within the narratives, ranging up to approximately 10,000 occurrences. The average mean for these terms typically fell within a range of 100 to 160, indicating their recurrent but varied usage across the narratives.

For instance, the term "patient" demonstrated an average occurrence of around 100 with a standard deviation of approximately 98, prevalent in nearly 10,000 texts. Similarly, terms like "treatment," "x," "tumor," "tests," and "prostate" showcased varying mean occurrences, averaging around 90 to 158, with respective standard deviations and occurrences in a substantial number of texts.

The graphical representation depicted the relationship between the total number of texts and the average size along with the standard deviation for the prominent medical terms analyzed. The analysis provides a clear understanding of the prevalence and variability in the usage of medical vocabulary within the narrative texts, offering valuable insights into the frequent occurrence and distribution of these medical terminologies.

G. WordCloud for Recommendation Word Usage:

The WordCloud visualization of doctor recommendations aimed to illuminate the most frequent and reiterated concepts within the text. The WordCloud, showcased in Figure 9, visually depicted the frequency of words within the doctor recommendations, emphasizing the most prominent terms in these advisories.

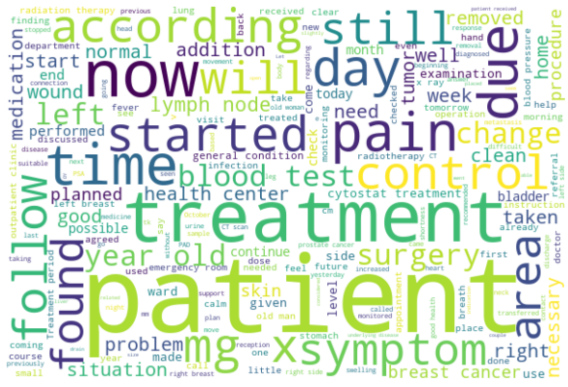


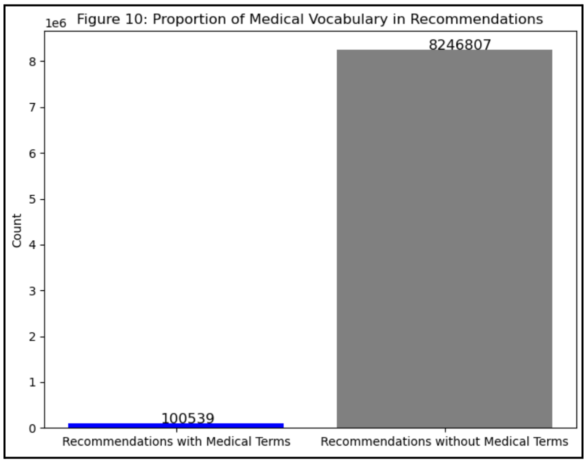
Figure 4: WordCloud representation of the Recommendations

Among the top words depicted in the WordCloud, recurrent terms such as "treatment," "patient," "pain," "symptoms," "time," "found," "follow," "started," "according," and "mg" surfaced with notable prominence. These terms indicate the recurrence of essential themes within the doctor recommendations, emphasizing the significance of specific aspects such as treatment, patient care, symptom management, and medication dosage.

The WordCloud offered a quick and intuitive insight into the emphasis of certain terms within the doctor recommendations, enabling a clearer understanding of the predominant advice conveyed in these narratives. This visualization helps to identify the most recurring and pivotal aspects addressed within the doctor recommendations.

H. UMLS for Medical Vocabulary Proportion:

The application of the Unified Medical Language System (UMLS) enabled the quantification of medical terminologies present within the doctor recommendations. Through this process, the team successfully mapped terms extracted from the recommendations to the extensive UMLS database, allowing an evaluation of the prevalence of medical vocabulary used within the doctor's advice. This methodology offered crucial insights into the prevalent medical language incorporated within the recommendations.



The analysis covered a significant volume of recommendations, totaling 8,347,346 instances. Within this dataset, 100,539 recommendations were identified as containing medical vocabulary, constituting approximately 1.20% of the entire set. This finding underscores the relatively limited use of explicit medical terminologies within the doctor recommendations.

The proportion of medical vocabulary in the recommendations, despite being around 1.20%, is still substantial in terms of the sheer volume of recommendations processed. This reveals the prominence of medical terminology in a proportion of the doctor recommendations, contributing to a better understanding of the integration of medical language within the narratives. Figure 10 graphically represents this analysis, offering a clear visual depiction of the prevalence of medical vocabulary within the doctor recommendations.

I. Empath Categorization of Recommendation Context:

The repeated use of Empath categorization allowed for a deeper exploration of prevalent themes, sentiments, and emotional nuances within the doctor's recommendations. This comprehensive analysis unveiled the dominant thematic elements, shedding light on the emotional context embedded within the recommendations. Figure 11 illustrates the Proportion of Categories in Doctor Recommendations Exceeding 0.003. Notably, the primary categories identified in the recommendations closely paralleled those discovered within the narrative texts, encompassing themes such as health, medical emergencies, healing, and pain. This consistency underscores the recurrent emphasis on essential topics like health and recovery, indicating their critical role in advising patients.

The thematic cohesion observed between the narratives and the doctor's recommendations highlights the recurring importance of key subjects, reflecting their consistent presence in advising patients. This meticulous examination of Empath categories within the doctor recommendations offers valuable insights into their thematic emphasis and emotional context. It underscores the persistent presence of critical themes and the emotional backdrop that plays a pivotal role in guiding patients through the medical advice they receive.

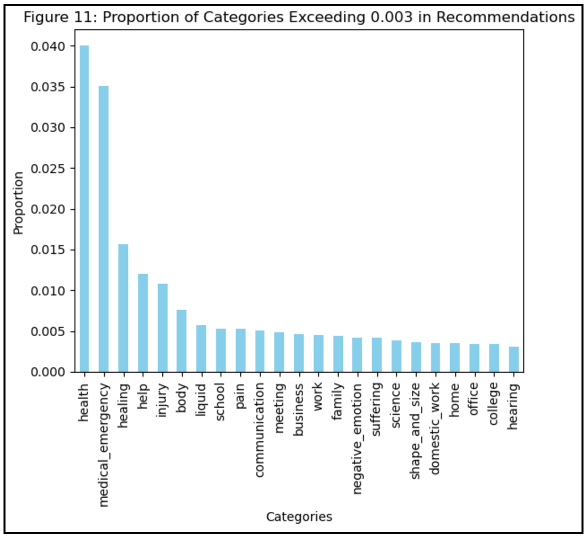


Figure 11: Proportion of Categories Exceeding 0.003 in Recommendations

J. Graphical User Interface (GUI):

Designed an intuitive graphical interface using Tkinter to facilitate the execution of various predefined tasks. The purpose of this interface is to streamline the evaluation process for external users or assessors. Through a user-friendly button interface, individuals can conveniently access the results and graphs generated from the analysis.

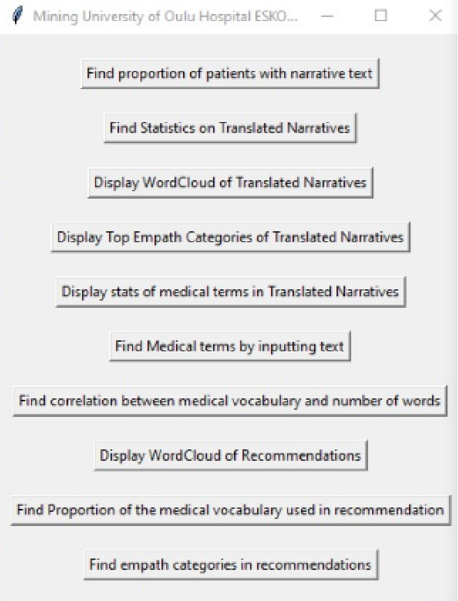

Mining University of Oulu Hospital ESKO...
Find proportion of patients with narrative text
Find Statistics on Translated Narratives
Display WordCloud of Translated Narratives
Display Top Empath Categories of Translated Narratives
Display stats of medical terms in Translated Narratives
Find Medical terms by inputting text
Find correlation between medical vocabulary and number of words
Display WordCloud of Recommendations
Find Proportion of the medical vocabulary used in recommendation
Find empath categories in recommendations

Figure 12: Graphical User Interface (GUI)

IV. Overall Discussions

In the examination of the overall findings and outcomes of this project, there is a significant convergence observed between the identified themes and emotional context prevalent in both narrative texts and doctor recommendations. The empath categorization method has offered valuable insights into the consistent thematic emphasis on health, medical emergencies, and healing, reaffirming their paramount importance in the advice given to patients. However, as a potential improvement for future studies, exploring alternative methods, such as employing UMLS for categorization or refining semantic types to enhance the extraction of medical terms, could be beneficial. Additionally, seeking or devising an automated approach for the filtration of medical terminologies might prove more efficient than the current manual curation method.

Moreover, the analysis highlighted the need for continued exploration into more precise methods for identifying medical terminologies. While the project offered a comprehensive methodology in extracting these terms, further enhancement in the precision of terminological extraction can be sought through the exploration of varied tools and strategies. This detailed examination marks a novel approach in comprehending the prevalent themes within the narratives and recommendations, contributing to a better understanding of the language, emotions, and thematic content embedded in medical advice. The improvement in automatic filtering and more refined categorization methods stands as a promising area for future research, offering potential advancements in the comprehension of medical narrative content.

V. Conclusion

This project's comprehensive analysis of medical narrative texts and doctor recommendations provided invaluable insights into the language, themes, and emotional undercurrents present in medical advice. Despite the complexity and depth of the tasks, challenges persisted in accurately discerning medical terminologies. The manual curation and semantic type refinement processes pointed to the need for more sophisticated, automated methods. The project honed skills in natural language processing, data analysis, and statistical inference, offering a deeper understanding of various tools and techniques in medical text analytics.

Looking ahead, there's a clear potential for refining terminological extraction and categorization. Automated approaches should be explored to streamline the filtering of non-medical terms and enhance the precision of semantic type assignment. Employing alternative tools, such as the Unified Medical Language System (UMLS), could offer more accurate categorization and improved extraction methods. These insights and skills gained have laid a robust foundation for future research, promising more refined and precise analyses in the domain of medical narrative texts. This paves the way for more sophisticated methodologies and deeper insights into the nuances of medical language.

VI. References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3270933/
2. https://link.springer.com/article/10.1007/s11042-022-13428-4
3. https://www.nlm.nih.gov/research/umls/about_umls.html
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9108578/
5. https://github.com/azizulkawser/Medical_text_analyzer

VII. APPENDICES

1. Appendix: Semantic Types - Tree Numbers and Term Unique Identifiers (TUIs)

Semantic Types Tree Numbers:

A1.1.2, A1.1.3.2, A1.1.4, A1.2.1, A1.2.2.1, A1.2.2.2, A1.2.3.1, A1.2.3.2, A1.2.3.3, A1.2.3.4, A1.2.3.5, A1.3.3, A1.4.1.1.1.1, A1.4.1.1.2, A1.4.1.1.3.2, A1.4.1.1.3.3, A1.4.1.1.3.4, A1.4.1.1.3.5, A1.4.1.1.3.6, A1.4.1.1.4, A1.4.1.1.5, A1.4.1.2.1.5, A1.4.1.2.1.7, A1.4.1.2.2, A1.4.1.2.3, A1.4.2, A1.4.3, A2.1.4.1, A2.1.5.1, A2.1.5.2, A2.1.5.3, A2.1.5.3.1, A2.1.5.3.2, A2.1.5.3.3, A2.2.1, A2.2.2, A2.3, A2.3.1, B1.1.1, B1.1.2, B1.2, B1.3.1, B1.3.1.1, B1.3.1.2, B1.3.1.3, B1.3.2.1, B2.2.1.1, B2.2.1.1.1, B2.2.1.1.1.1, B2.2.1.1.2, B2.2.1.1.3, B2.2.1.1.4, B2.2.1.1.4.1, B2.2.1.2, B2.2.1.2.1, B2.2.1.2.1.1, B2.2.1.2.1.2, B2.2.1.2.2, B2.2.1.2.3, B2.3, R3.1.2, R5.6

Semantic Types Term Unique Identifiers (TUIs):

T007, T004, T005, T018, T019, T020, T023, T024, T025, T026, T028, T200, T121, T122, T125, T126, T127, T129, T192, T130, T131, T114, T116, T197, T196, T031, T168, T022, T030, T029, T085, T086, T087, T088, T034, T184, T032, T201, T054, T055, T056, T058, T059, T060, T061, T063, T039, T040, T041, T042, T043, T044, T045, T046, T047, T048, T191, T049, T050, T037, T154, T163

This appendix provides the semantic types along with their corresponding tree numbers and term unique identifiers (TUIs) used during the filtering process for the extraction of medical terminologies within the narrative texts.