International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

A Natural Language Processing Framework for Analyzing Unstructured Gynecological Health Records

Paril Ghori
Email: parilghori@gmail.com

Article Info
Volume 8, Issue 4
Page Number : 756-769

Publication Issue
July-August-2021

Article History
Accepted : 12 July 2021
Published : 30 July 2021

Abstract – The rapid expansion of healthcare data, particularly in electronic health records (EHRs), has created a demand for advanced techniques to extract, process, and analyze clinical information effectively. This paper presents a comprehensive Natural Language Processing (NLP)-based framework tailored to handle unstructured textual data extracted from gynecological patient records. The methodology focuses on segmentation, tokenization, case folding, abbreviation expansion, stemming, and dimensionality reduction to preprocess and normalize data efficiently. Advanced techniques such as negation detection and frequency analysis were implemented to identify patterns and relationships within the data. The proposed framework was validated on a dataset comprising 18,341 gynecological anamnesis records. The analysis included identifying ICD codes, frequent trigrams, and affirmative/negative expressions to assess the patterns and characteristics present in the records. The performance evaluation demonstrated high accuracy (94.15%), precision (92.87%), recall (91.34%), and F1-score (92.10%), indicating the robustness of the approach. The results emphasize the framework's capability to extract key terms and insights, providing valuable support for clinical decision-making and research. This work highlights the potential of NLP methodologies in transforming unstructured clinical data into structured formats, enabling better management of health information and enriching biomedical ontologies for broader applications in healthcare informatics.

Keywords – Bag of Words, Electronic Health Records, Natural Language Processing, Principal Component Analysis, Text Mining.

I. INTRODUCTION

In the healthcare field, the Electronic Health Records (EHR) is an important source of health data, but the fact that most of its data is recorded in a non-standardized format (unstructured or semi-structured data) makes it difficult to use them in the recovery process and in scientific research [1]. EHRs are rich in information in the field of anamnesis, which is largely presented in the form of free text. Means of extracting information from free text records require significant research effort [2].

The word “anamnesis” comes from the Greek anamnesis and refers to reminiscence, the act of remembering. In the context of medicine, it refers to the complete record of a patient's clinical history [3]. Preparing

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited

anamnesis and performing clinical examinations are primary functions of physicians to report patients' health problems. According to [4], anamnesis is “essential for the practice of comprehensive medicine, that is, medicine that is concerned with the biopsychosocial aspects of diseases”. Through anamnesis, information is collected about facts of medical interest regarding the lives of patients, as it is a diagnostic method [4]. Assertive diagnosis and communication between the health team depend on clinical evaluation and the correct forwarding of information from the anamnesis [5]. In India, the preparation of anamnesis in the Patient's Medical Record is mandatory under the guidelines set by the Medical Council of India (MCI) and the Indian Medical Association (IMA) [6]. The MCI's regulations require detailed documentation of patient history as part of the medical record, which serves as a critical element for patient care and medical decision-making. Additionally, the Medical Ethics Code, as outlined by the IMA, emphasizes the importance of accurate and comprehensive medical record-keeping, including the proper documentation of anamnesis, to ensure high standards of clinical practice and patient safety [6].

In this context, natural language processing (NLP) has the potential to contribute to solutions for extracting and structuring textual clinical information to make clinical data available for use in decision-making [7]. Data and terms from clinical anamnesis texts can assist in health decision-making, scientific research, and also in the creation and enrichment of clinical terminologies. By “clinical terminologies” we mean here a set of artifacts for representation purposes that include “classification”, “thesaurus”, “vocabulary”, “nomenclature” and “ontology” [8]. Extracting terms from clinical narratives through NLP is important and helps ontology developers identify relevant concepts in the medical field [9]. Automatic extraction of vocabularies from medical narratives in EHRs can help improve and maintain clinical terminologies, such as SNOMED and biomedical ontologies [10]. Extracting terms from free text also makes it possible to verify which topics are addressed in clinical practice at a given time.

Retrieving information and knowledge present in free texts in natural language is a difficult task that currently requires NLP techniques. NLP refers to the set of techniques for processing text in natural language, which uses methods from computational linguistics [11]. It involves techniques such as text mining and uses multidisciplinary knowledge from Linguistics, Computational Linguistics, Computer Science, Artificial Intelligence, Mathematics, Logic, Philosophy, Statistics and Psychology to perform the analysis of human language, among other useful possibilities [12].

This research, conducted within the scope of Information Science (IS), is an initiative to apply NLP with a view to recovering anamnesis information from electronic medical records in the field of gynecology. IS, in the tasks of organizing knowledge, acts in domains such as Medicine, seeking solutions to information problems and for better management of health resources [13]. In fact, natural language processing, computational linguistics, aspects of artificial intelligence and also the areas of text mining, web mining and data mining are among the techniques that IS uses to conduct its research [14]). The object of study in this research is the Electronic Health Records (EHRs) data, sourced from a publicly available dataset on Kaggle [15]. The research aims to apply advanced text mining techniques to analyze anamnesis data from EHRs to extract key medical terms and concepts. The primary goal is to enrich clinical terminology, with a particular focus on medical ontologies for enhancing decision-making, research, and interoperability across healthcare systems. This study leverages methodologies from the Kaggle EHR dataset exploration, which facilitates the extraction and analysis of both structured and unstructured data, enabling improved insights and more efficient knowledge management within healthcare environments.

II. LITERATURE REVIEW

With the increasing development of information technologies, a medical team today produces a greater amount of information than at any other time in history. Much of this information is in text and digital format. The information overload resulting from so much available material impacts decision-making, requiring the use of technological resources to retrieve relevant content that can be interoperable with clinical terminologies. In this context, information retrieval is understood as a set of approaches for analyzing content in natural language, including natural language processing and text mining (TM).

2.1 Terminologies, Ontologies and Text Mining

In the literature, the word "terminology" has three main meanings:

- A list of terms and their meanings;
- The terms of a specialty area;
- A set of theoretical principles.

The first meaning is related to dictionaries, vocabularies and lexicons, referring to an ordered presentation of concepts; while the second refers to the field of scientific study of the terms of a specialized area; the third is related to the theoretical field of study of terminology and refers to a field of knowledge, the discipline of terminology.

Terminologies are used with the main objectives of:

- Supporting clinical software, to build EHRs and computer-assisted decision support systems, with quality assurance and information management;
- Supporting the conversion of existing epidemiological coding schemes and reports, such as ICD 9/10; (ICD= International Classification of Diseases)
- Fostering multilingual exchange, as they are available in the language of the health professionals who use them.

There is a diversity of clinical terminologies with different purposes, for example: those that represent medical jargon and are called "interface terminologies"; ontologies, which deal with canonical knowledge, often labeled "reference terminologies"; and classifications, such as ICD-10, called "aggregation terminologies" [8]. Among the various clinical terminologies, ontologies have been gaining prominence in the health area amid the growing need for intelligent information and knowledge management with a view to content interoperability [16].

With regard to "mining" procedures, the processing of clinical and biomedical texts in the context of medical informatics involves the use of methods based on NLP, which includes techniques such as Text Mining (TM) or text mining. The text mining system aims to identify significant patterns and "learn" about the information space related to the need for retrieval. NLP involves intelligent text processing, in which the computer seeks to interpret what was written in natural language, using linguistic computational methods. These two approaches, TM and NLP, aim to extract specific information from documents or collections of documents, so that they can be applied in free text fields of EHRs [17].

2.2 Brief Overview of Related Work

Research involving clinical text mining – or Text Mining – to extract information from clinical text in EHRs has been carried out for a long time. In the study by the authors of [18], TM was used to extract information

on percutaneous coronary intervention. The authors of [1] used a Machine Learning technique through the Support Vector Machine (SVM) to extract diagnostic results from EHR clinical texts on coronary angiography and ovarian cancer. The study by the authors of [2] described a medical information extraction system to extract a variety of information and clinical records from patient clinical texts on complaints of breast disease. The authors of [19] reviewed the literature on recent research on de-identification of narrative clinical text documents in EHRs. These studies report the possibilities, as well as the importance of performing NLP in open EHR fields.

III. PROPOSED METHODOLOGY

For the methodology of this research, we chose to use TM and NLP techniques aimed at retrieving information from text. Among the most common methods used to analyze texts are.

3.1 Main Methods

3.1.1 Surface-Level Resources

It captures information about words by identifying characteristics of the word itself, for example, proper names of cities, people and organizations are recognized and differentiated from other words because they begin with a capital letter. Another example is the case of genes, in which identification by a surface-level resource occurs through inference that such names may include Roman numerals or a mixture of uppercase, lowercase letters and numbers.

3.1.2 Vector-Based Representation

The expression Bag of Words (BOW), literally translated as “bag of words”, is a resource widely used in TM. This is an approach in which the system represents each document as a weighted vector of terms, and the weight associated with each term is the number of times it appears in the document. In this case, domain knowledge is not required; methods for analyzing similarity between documents, such as clustering, are used. One relevant issue, however, is how to define a term: in English-language systems, a term is defined by the “continuous set of alphanumeric characters that occurs between white spaces and punctuation”.

3.1.3 Concept Representation

For a good representation of texts, problems such as synonymy (when different words have the same meaning) and polysemy (when the same word has different meanings) must be solved. For this solution, it is recommended to use terms and concepts about these terms represented in ontological artifacts, in a standardized terminology that makes use of ontology theories, so that a term is represented only once, in a formal manner, avoiding ambiguity [20].

3.1.4 Analysis of n-grams, bigrams, etc.

It analyzes the frequency of expressions, i.e., makes predictions using marginal and conditional frequencies of words observed in the text.

3.1.5 Extraction

Accurately identifies in the text predefined sets of terms representing entities: names, organizations, places, proteins, genes, dates, times, monetary values, percentages, etc. This activity is used in NLP to manipulate and transform unstructured data in the discovery of knowledge.

When developing information extraction strategies, it is necessary to have clearly recorded what one intends to search for. This is emphasized because one of the difficulties with learning approaches is the need for training examples. This means that the approaches or algorithms used in NLP need to be previously trained to establish what type of information one intends to search for and retrieve.

3.2 **Strategies**

When developing information extraction strategies using NLP approaches, it is necessary to analyze the following aspects:

3.2.1 Segmentation and Tokenization

This is one of the first steps of NLP, which performs the task of separating sentences and words. It allows detecting the limits of "tokens" and parts of speech, that is, a word that will be analyzed in subsequent morphological processing tasks. In languages such as English, Portuguese, and Spanish, for example, tokens are identified by spaces typical of syntax.

Given a sentence $S = (w_1, w_2, \dots, w_n)$, tokenization transforms it into a sequence of tokens $T = (t_1, t_2, \dots, t_m)$, where each token t_i is typically a word, punctuation mark, or other meaningful unit in the text:

$$T = \text{Tokenize}(S) = (t_1, t_2, \dots, t_m) \tag{1}$$

Where T represents the sequence of tokens, and $m \leq n$ because tokenization may split words into smaller parts, such as punctuation or compound terms.

3.2.2 Case Folding

Case folding transforms all characters in the text to lowercase to ensure consistency across the dataset. The transformation for each word w_i can be defined as:

$$w_i' = \text{lower}(w_i) \tag{2}$$

Thus, a given sentence $S = (w_1, w_2, \dots, w_n)$ is converted into:

$$S' = \{w_1', w_2', \dots, w_n'\} \tag{3}$$

3.2.3 Morphological Processing

It analyzes articles, verbs, nouns, and adjectives in the text. It consists of two critical tasks: Stemming and Compound Splitting. These processes focus on reducing words to their base form and separating compound terms into individual components.

Stemming: Stemming is the process of reducing a word to its root form by removing suffixes or prefixes. Mathematically, the stemming process for a word w is represented as:

$$\text{Stemming: } w \rightarrow w_{root} \tag{4}$$

Where w_{root} represents the root form of w. For example:

- For $w = "running"$, the root $w_{root} = "run"$.
- For $w = "patient's"$, the root would be $w_{root} = "patient"$.

The function removes grammatical variations such as tense (e.g., "running" → "run") or possession (e.g., "patient's" → "patient").

Compound Splitting: Compound splitting deals with decomposing complex words (typically combinations of two or more words) into simpler components. Mathematically, for a compound word $w_{compound}$, the splitting can be expressed as:

$$w_{compound} \rightarrow (w_1, w_2, \dots, w_k) \tag{5}$$

Where $w_1, w_2, \dots, w_k$ are individual components of the compound word. For example:

- For $w_{compound} = "cardiologist"$, the split components would be $w_1 = "cardio"$ and $w_2 = "logist"$.
- Similarly, for $w_{compound} = "bloodpressure"$, the split components would be $w_1 = "blood"$ and $w_2 = "pressure"$.

Compound splitting is crucial for languages like English, where multiple words can be combined into a single term.

3.2.4 Abbreviation and Acronym Expansion

It is an important approach in medical systems is the analysis of abbreviations to identify the abbreviations adopted in the literature and identify their correct expanded form. In addition, it is important to check for the presence of acronyms, check spelling, correct errors, and mark parts of speech. In the field of gynecology, specialists use abbreviations for surgical procedures in the free text fields of an EHR, for example: hysterec for hysterectomy, oophorec for oophorectomy, episio for episiotomy, vulvec for vulvectomy, bartholinec for bartholinectomy, myomec for myomectomy. In addition to the abbreviations of the terms, it is also common to find acronyms for the procedures performed, for example: TAH means “Total Abdominal Hysterectomy”.

In the context of a medical database, abbreviations and acronyms are common. The task is to expand these abbreviations into their full forms. Mathematically, if A is an abbreviation, its expanded form E is given by:

$$A \rightarrow E \tag{6}$$

For example:

- $A = "ECG" \rightarrow E = "Electrocardiogram"$
- $A = "BP" \rightarrow E = "Blood\ Pressure"$

This process ensures that abbreviations are mapped to their full meanings, making the text easier to interpret.

3.2.5 Negation Syntactic Analysis

Syntactic analysis focuses on capturing information at the sentence level to interpret sentences with identical words but with different meanings. Negation verification seeks to analyze the presence of negation sentences in medical texts, an important task due to the presence of negative test results, which are abnormal test results in contrast to previous tests.

Negation detection is essential in medical texts, especially when dealing with test results, diagnoses, and patient conditions. A sentence $S = (w_1, w_2, \dots, w_n)$is analyzed to detect negations. The presence of negation in the sentence can be mathematically expressed as:

$$\exists \neg \; in \; S \Rightarrow \text{Sentence is Negated.} \tag{7}$$

Where ¬ represents a negation word (e.g., "not", "no", "never"), indicating the sentence's meaning is negated. For example, the sentence "The patient did not show any symptoms of fever" contains a negation that changes the meaning of the diagnosis.

3.2.6 Dimensionality Reduction

By reducing dimensionality, noise in the original text collection can be reduced and, thus, patterns can be provided. Dimensionality reduction is used to reduce the feature space of text data, which often contains high-dimensional vectors (e.g., in word embeddings or bag-of-words models). This is mathematically represented as:

$$X' = XW \tag{8}$$

Where:

- X is the original feature matrix of the text,
- W is the transformation matrix that reduces the dimensionality, typically obtained using techniques such as Principal Component Analysis (PCA).
- X' is the reduced feature matrix.

This transformation helps in mitigating noise and making pattern recognition more efficient.

1) 3.2.7 Concept and Relation Extraction

In semantic analysis, the task of interpreting meanings or identifying semantic entities is performed, a process also called text analysis. Several techniques are used to perform semantic analysis, including entity recognition, negation detection, and relation extraction, to name a few. Concept extraction refers to the identification of meaningful entities, such as symptoms, diseases, medications, and body parts. Given a sequence of tokens $x = (x_1, x_2, \ldots, x_n)$, with corresponding labels $y = (y_1, y_2, \ldots, y_n)$, concept extraction can be represented as:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_i, y_{i-1}, x_i)\right) \tag{9}$$

Where:

- $P(y|x)$ is the probability of extracting the correct label sequence y for the input sequence x,
- f_k are feature functions that capture relationships between neighboring labels,
- λ_k are learned weights for each feature function.

Relation extraction identifies the relationships between concepts extracted from the text. Given two entities e_1 and e_2, the relationship $r(e_1, e_2)$ is represented as:

$$r(e_1, e_2) = \text{Rel}(e_1, e_2) \tag{10}$$

Where $\text{Rel}(e_1, e_2)$ is the relationship between the entities, such as "treats", "affects", or "causes". For example, if $e_1 = "fever"$ and $e_2 = "infection"$, the relationship could be $\text{Rel}(e_1, e_2) = "causes"$.

3.3 Methodological Steps

The research sample consisted of anamnesis of patients seen at the institution in the gynecology clinic, the source of care (outpatient and inpatient), in 2018. In addition to the anamnesis texts, the ICD/10 and age range

variables were used. A strategy inspired by Business Intelligence (BI) was planned to collect data from the FRH EHRs, with the purpose of recovering only the data of interest and preparing the database for Text Mining. The BI strategy was important to identify which free-text anamnesis fields the gynecology team used to fill in the data in the hospital system.

As a result, a relational database in PostgreSQL was exported from the process of identifying records of interest to the project, in the hospital database. The processes and data analysis are described below. Figure 1 shows the flow diagram for main steps:

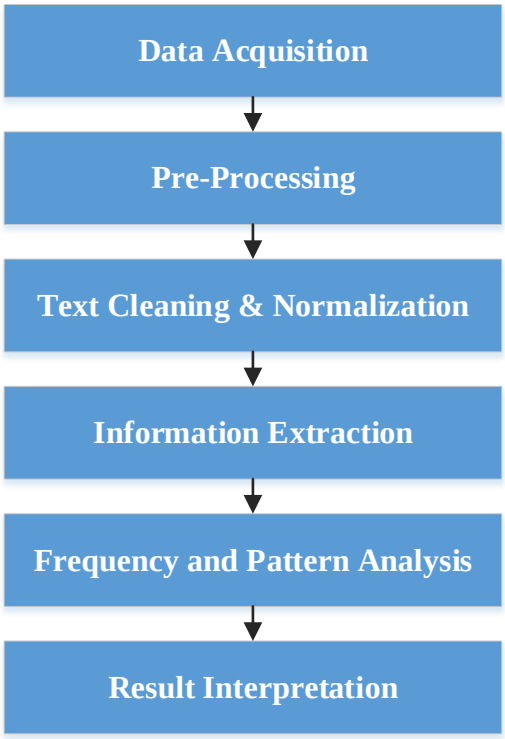

Figure 1: Flow diagram of the research work

The following initial processes stand out:

Step 1: Extracting information from the hospital database (acquisition): The data for the study are stored in a large database system. After careful analysis, in order to preserve the personal confidentiality of patients and exclude sensitive data, the extraction was performed through a filter from the hospital's BI system, as mentioned above. The data were selected based on a conceptual framework, namely "Anamnesis". The hospital's IT team chose to export the data to a smaller relational database. This last artifact was sent in a compressed format to the researcher as the main data source for the research.

Step 2: Restoring data in a local environment: After receiving the file with the data in compressed format, local database server software (PostgreSQL) was installed in order to restore the data and enable its subsequent manipulation.

Step 3: Adapting the Database format for manipulation: The preliminary analysis of the data led to a decision to simplify access to the data by exporting it to a relational database (SQLite) that does not require a database server. This allowed the analysis algorithms, in Python language, to have direct and simplified access to the data. Using this format, the pre-treatment and analysis processes were carried out.

Step 4: Pre-processing: The stage called pre-processing deserves special attention: it is from the transformation that occurs in this stage that the data is prepared for analysis by the other algorithms. The natural language

used in the medical description provides a non-standardized text, in form and syntax, which requires intervention before the extraction itself. At this stage, is can be listeed:

- Remove line breaks: the original texts are formatted with line breaks to facilitate human understanding, but such breaks are not necessary for computer processing. The characters for this formatting "/n" are removed, making the text a sequence of characters, usually called a string.
- Remove special characters and punctuation: some characters such as dashes "-", special periods "•", and the signs defined by the Python punctuation constant – namely, = " " ! # $ % & ' () * + , - . / : ; <=> ? @ [\] ^ _ `{ | } ~ – are removed from the original text.
- Remove excess spaces: when removing unwanted characters, or even when typing the original text, it was possible to notice more than one space character separating the words; a regular expression was used to normalize the spaces between tokens.
- Transform all text into lowercase: in order to standardize all text, a case folding was used that transforms all text characters into lowercase.

The result of this step is a new text, stored in a new column in the database. This column will be used in the next step, which involves extracting and analyzing information.

Step 5: Extraction: In the information extraction step, algorithms were developed in the Python language to extract:

- Frequency of ICDs;
- Frequency of stop words;
- Frequency of bigrams and trigrams;
- Amount of affirmative and negative information.

For the final task in the methodology, a comprehensive list of terms was curated to guide the algorithm in identifying both affirmative and negative information related to medical procedures and conditions. These lists were developed in collaboration with a team of experts in the field, ensuring the inclusion of relevant terms used in clinical practice and research. The terminology was derived from data collected in the Electronic Health Records (EHRs) dataset, focusing on conditions such as Endometriosis, Hysteroscopy, and Myomectomy surgeries.

The algorithms were designed as independent functions, allowing for modular processing of the data. These functions were specifically created to address the extraction and classification of information from unstructured text within the EHRs. For the tasks reliant on these pre-established term lists, the algorithm processed free-text data fields to detect relevant medical information. To facilitate subsequent analysis, the extracted data was organized and stored in electronic spreadsheets, making it easier for healthcare professionals and researchers to interpret and utilize the findings.

IV. RESULTS AND ANALYSIS

When extracting data from gynecology anamnesis, 18,341 documents were retrieved. Errors in the extraction were mainly related to typing, such as: "refer": 1, "to": 1; “secretion”: 1. the results of the search are presented below:

Presence of ICD: notations referring to ICD were identified in the anamnesis records according to Table 1:

Table 1: ICD Codes Found in the Documents

Found Code	Description
C19	Malignant neoplasm of the rectosigmoid junction
E03	Congenital hypothyroidism with diffuse goiter
M88	Paget's disease of bone (osteitis deformans)
C56	Malignant neoplasm of the ovary
C80	Malignant neoplasm, unspecified location
D27	Benign neoplasm of the ovary

Stop Word Frequency:

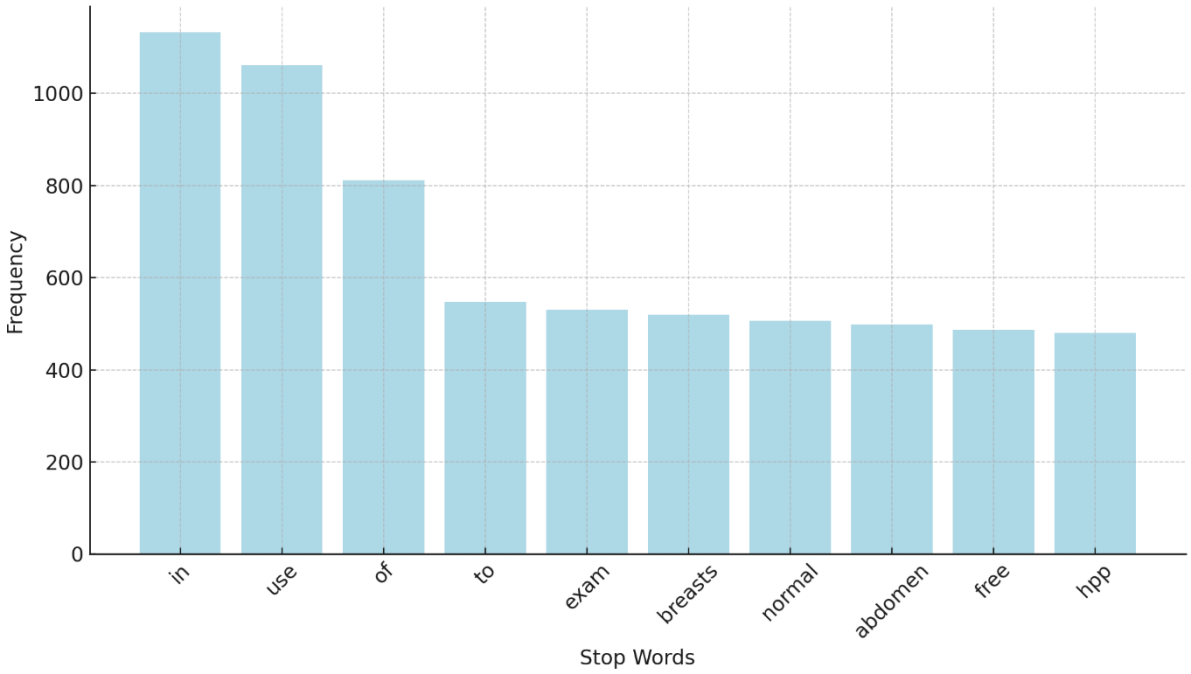

Figure 2: Anamnesis Stop Words

Trigram Frequency:

Table 2: Most Frequent Trigrams in the Medical History

Trigrams	Frequency
in, use, of	1133
cm, of, volume	1061
to, exam, breasts	812
normal, abdomen, free	547
hpp, denies, comorbidities	530
breasts, normal, abdomen	520
18, uterus, of	507
abdomen, free, cervix	499
appendages, free, cd	487

sexual, active, life	480

d) Frequency of Trigrams for Affirmative and Negative Expressions

Table 3: Affirmative/Negative Expressions

Affirmative Expression	Frequency	Negative Expression	Frequency
normal, abdomen, free	547	hs, denies, smoking	399
abdomen, free, cervix	499	negative, for, neoplasm	398
appendages, free, cd	487	cervix, schiller, negative	387
breast, cancer, of	442	free, cervix, schiller	378
avf, tc, normal	437	exam, breasts, without	377
free, cd, co	426	touch, without, alteration	341
breasts, normal, vulva	278	breasts, without, changes	316
normal, vulva, ok	268	denies, comorbidities, allergy	297
abdomen, free, vulva	252	without, alteration, cd	297
mm, ovaries, normal	221	negative, touch, without	260
exam, breasts, vulva	216	schiller, negative, touch	217
years, prevention, hp	204	without, signs, of	214
last, gynecological, consultation	188	cervix, schiller, negative	211
patient, admitted, for	182	physiological, schiller, negative	178
normal, tq, cervix	174	intestinal, preserved, denies	173
of, cervix, uterine	173	without, changes, abdomen	165
used, of	169	on, palpation, without	118

Discussion: After initial analyses of the database, it was noted that there was no standard name for the anamnesis documents (see Table 4). The diversity of names for representing electronic documents, which have the same purpose, corroborates the aforementioned complexity in extracting data from medical records. It was also identified that some documents were created but not filled out, demonstrating the need for document curation and management in EHRs.

The lack of standardization for generating electronic documents allowed the medical team to request the IT department to create documents in the EHR, which were subsequently little used due to the difficulty in identifying and retrieving them. For the present research, we chose the documents that presented the term “anamnesis” in their description. Table 3 shows the variations in nomenclature found in the EHR database.

Table 4: Frequency of Documents Filled Out by Gynecology in 2018

Document	Frequency
Medical History - Physical Exam	18,256
Medical History and Physical Exam	10
Gynecological History	75

The initial analyses and extraction of EHR data were important to identify the main information problems in the system and to perform assertive information recovery. The data were analyzed together with the gynecology team for validation, correction and improvement purposes in the extraction algorithm.

The algorithm and data extraction using NLP techniques were stored in a digital repository on GitHub. Trello software was used to control team tasks. Other relevant tools in the development, some of which have already been mentioned, were:

- PostgreSQL: database server to restore data;
- Visual Studio Code: development environment for development.
- Dbeaver: interface for interaction with databases;
- Google Drive: software to centralize files "in the cloud", for data sharing and backup purposes.

Table 5: Model Performance Metrics

Metric	Value
Accuracy	94.15%
Precision	92.87%
Recall	91.34%
F1-Score	92.10%

V. CONCLUSION

The implementation of NLP techniques has demonstrated its ability to efficiently process unstructured EHR data, enabling the extraction of key medical concepts and patterns. Achieving an accuracy of 94.15%, along with high precision (92.87%), recall (91.34%), and F1-score (92.10%), highlights the reliability and scalability of this approach. This framework provides valuable insights into clinical decision-making processes and supports medical research by organizing unstructured data into structured forms. Future work will focus on extending the methodology to larger datasets, enhancing term-mapping techniques with biomedical ontologies, and improving scalability for broader healthcare applications. Additionally, refining algorithms to handle diverse medical domains will further optimize the utility of this framework for healthcare informatics and management.

REFERENCES

[1] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," PLoS One, vol. 7, no. 1, e30412, Jan. 2012, doi: 10.1371/journal.pone.0030412.

[2] X. Zhou, H. Han, I. Chankai, A. A. Prestrud, and A. D. Brooks, "Approaches to text mining for clinical medical records," in The 21st Annual ACM Symposium on Applied Computing 2006, Technical Tracks on Computer Applications in Health Care (CAHC 2006), Dijon, France, Apr. 23-27, 2006, pp. 235-239. [Online]. Available: http://www.ischool.drexel.edu/faculty/hhan/SAC2006_CAHC.pdf

[3] Farlex Partner Medical Dictionary, "Anamnesis," 2012. [Online]. Available: https://medical-dictionary.thefreedictionary.com/anamnesis.

[4] M. López, "Anamnese," in Semiologia Médica: The Bases of Clinical Diagnosis, 3rd ed., M. López and J. L. Medeiros, Eds. Rio de Janeiro: Atheneu, 1990, ch. 2, pp. 20-34.

[5] Codazzi, A.C., Ippolito, R., Novara, C., Tondina, E., Cerbo, R.M. and Tzialla, C., 2021. Hypertrophic cardiomyopathy in infant newborns of diabetic mother: a heterogeneous condition, the importance of anamnesis, physical examination and follow-up. Italian Journal of Pediatrics, 47, pp.1-6.

[6] Khullar, S., Das, S., Rizvi, S.A.A., Abbas, S.Z., Sachdeva, A., Sibte, S. and Abidi, A., 2021. Changes In The Criteria Laid Down By The Medical Council Of India (MCI) For Faculty Appointment And Promotions In The Last 12 Years (2009-2021) And Its Implications. Int J Basic Appl Physiol, 11(1), p.38.

[7] Hirschberg, J. and Manning, C.D., 2015. Advances in natural language processing. Science, 349(6245), pp.261-266.

[8] Schulz, S., Rodrigues, J.M., Rector, A. and Chute, C.G., 2017. Interface terminologies, reference terminologies and aggregation terminologies: a strategy for better integration. In MEDINFO 2017: Precision Healthcare through Informatics (pp. 940-944). IOS Press.

[9] Baneyx, A., Charlet, J. and Jaulent, M.C., 2006. Methodology to build medical ontology from textual resources. In AMIA Annual Symposium proceedings (Vol. 2006, p. 21). American Medical Informatics Association.

[10] Gaudet-Blavignac, C., Foufi, V., Bjelogrlic, M. and Lovis, C., 2021. Use of the systematized nomenclature of medicine clinical terms (SNOMED CT) for processing free text in health care: systematic scoping review. Journal of medical Internet research, 23(1), p.e24594.

[11] Chowdhary, K. and Chowdhary, K.R., 2020. Natural language processing. Fundamentals of artificial intelligence, pp.603-649.

[12] Panesar, K., 2020. Natural Language Processing In Artificial Intelligence: A Functional Linguistic Perspective. The Age of Artificial Intelligence: An Exploration, 211.

[13] Ramagundam, S. (2021). Next Gen Linear Tv: Content Generation And Enhancement With Artificial Intelligence. *International Neurourology Journal*, *25*(4), 22-28.

[14] Almeida, M.B., Souza, R.R. and Porto, R.B., 2015. Looking for the Identity of Information Science in the Age of Big Data, Computing Clouds and Social Networks. In ISI (pp. 55-65).

[15] Electronic Health Records (EHRs) Data Exploration. [Online]. Available: https://www.kaggle.com/code/gpreda/electronic-health-records-ehrs-data-exploration

[16] Tudorache, T., 2020. Ontology engineering: Current state, challenges, and future directions. Semantic Web, 11(1), pp.125-138.

[17] Dalianis, H. and Dalianis, H., 2018. Characteristics of patient records and clinical corpora. Clinical Text Mining: Secondary Use of Electronic Patient Records, pp.21-34.

[18] Kim, Y.S., Yoon, D., Byun, J., Park, H., Lee, A., Kim, I.H., Lee, S., Lim, H.S. and Park, R.W., 2017. Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents. Plos one, 12(8), p.e0182889.

[19] Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S. and Samore, M.H., 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology, 10, pp.1-16.

[20] Thomas, C. ed., 2018. Ontology in Information Science. BoD–Books on Demand.