Research, Society and Development, v. 11, n. 6, e37711629184, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i6.29184

Applying Text Mining and Natural Language Processing to Electronic Medical Records for extracting and transforming texts into structured data

Aplicação de Mineração de Texto e Processamento de Linguagem Natural a Prontuários Médicos Eletrônicos para extração e transformação de textos em dados estruturados

Aplicación de Minería de Texto y Procesamiento de Lenguaje Natural a Registros Médicos Electrónicos para extraer y transformar textos en datos estructurados

Received: 04/13/2022 | Reviewed: 04/22/2022 | Accept: 04/28/2022 | Published: 04/30/2022

Diego Henrique Pegado Benício
ORCID: https://orcid.org/0000-0003-2750-0083
Digital Metropolis Institute, Brazil
Federal University of Rio Grande do Norte, Brazil
E-mail: diehpb@gmail.com
João Carlos Xavier Júnior
ORCID: https://orcid.org/0000-0003-1517-2211
Digital Metropolis Institute, Brazil
Federal University of Rio Grande do Norte, Brazil
E-mail: jcxavier@imd.ufrn.br
Kairon Ramon Sabino de Paiva
ORCID: https://orcid.org/0000-0001-9772-5101
Onofre Lopes University Hospital, Brazil
Federal University of Rio Grande do Norte, Brazil
E-mail: kaironpaiva@gmail.com
Juliana Dantas de Araújo Santos Camargo
ORCID: https://orcid.org/0000-0001-8692-5706
Maternity Hospital-School Januario Cicco, Brazil
Federal University of Rio Grande do Norte, Brazil
E-mail: juliana.camargo@ebserh.gov.br

Abstract

The recording of patients' data in electronic patient records (EPRs) by healthcare providers is usually performed in free text fields, allowing different ways of describing that type of information (e.g., abbreviation, terminology, etc.). In scenarios like that, retrieving data from such source (text) by using SQL (Structured Query Language) queries becomes an unfeasible issue. Based on this fact, we present in this paper a tool for extracting comprehensible and standardized patients' data from unstructured data which applies Text Mining and Natural Language Processing techniques. Our main goal is to carry out an automatic process of extracting, clearing and structuring data obtained from EPRs belonging to pregnant patients from the Januario Cicco maternity hospital located in Natal - Brazil. 3,000 EPRs written in Portuguese from 2016 e 2020 were used in our comparison analysis between data manually retrieved by health professionals (e.g., doctors and nurses) and data retrieved by our tool. Moreover, we applied the Kruskal-Wallis statistical test in order to statically evaluate the obtained results between manual and automatic processes. Finally, the statistical results have showed that there was no statistical difference between the retrieval processes. In this sense, the final results were considerably promising.

Keywords: Text Mining; Natural Language Processing; Electronic Medical Record; Anamnesis.

Resumo

O registro dos dados dos pacientes em prontuários eletrônicos (EPRs) pelos profissionais de saúde geralmente é realizado em campos de texto livre, permitindo diferentes formas de descrever esse tipo de informação (por exemplo, abreviatura, terminologia etc.). Em cenários como esse, recuperar dados de tal fonte (texto) usando consultas SQL (Structured Query Language) torna-se um problema inviável. Com base neste fato, apresentamos neste artigo uma ferramenta para extração de dados compreensíveis e padronizados de pacientes a partir de dados não estruturados que aplica técnicas de Mineração de Texto e Processamento de Linguagem Natural. Nosso principal objetivo é realizar um processo automático de extração, limpeza e estruturação de dados obtidos de PEPs de gestantes da maternidade Januário Cicco localizada em Natal - Brasil. Em nossa análise de comparação entre dados recuperados manualmente por profissionais de saúde (por exemplo, médicos e enfermeiros) e dados recuperados por nossa ferramenta foram usados

3.000 EPRs escritos em português. Além disso, aplicamos o teste estatístico de Kruskal-Wallis para avaliar estaticamente os resultados obtidos entre processos manuais e automáticos. Por fim, os resultados estatísticos mostraram que não houve diferença estatística entre os processos de recuperação. Nesse sentido, os resultados foram consideravelmente promissores.

Palavras-chave: Mineração de Texto; Processamento de Linguagem Natural; Prontuário Eletrônico; Anamnese.

Resumen

El registro de los datos de los pacientes en las historias clínicas electrónicas (HPE) por parte de los profesionales sanitarios suele realizarse en campos de texto libre, lo que permite diferentes formas de describir este tipo de información (p. ej., abreviatura, terminología, etc.). En escenarios como este, la recuperación de datos de dicha fuente (texto) mediante consultas SQL (Lenguaje de consulta estructurado) se convierte en un problema inviable. En base a este hecho, presentamos en este artículo una herramienta para extraer datos comprensibles y estandarizados de pacientes a partir de datos no estructurados que aplica técnicas de Minería de Texto y Procesamiento de Lenguaje Natural. Nuestro principal objetivo es realizar un proceso automático de extracción, limpieza y estructuración de datos obtenidos de PEP de gestantes en la maternidad Januário Cicco ubicada en Natal - Brasil. En nuestro análisis que compara los datos recuperados manualmente por profesionales de la salud (p. ej., médicos y enfermeras) y los datos recuperados por nuestra herramienta, se utilizaron 3000 EPR escritos en portugués. Además, aplicamos la prueba estadística de Kruskal-Wallis para evaluar estáticamente los resultados obtenidos entre procesos manuales y automáticos. Finalmente, los resultados estadísticos mostraron que no hubo diferencia estadística entre los procesos de recuperación. En este sentido, los resultados fueron considerablemente prometedores.

Palabras clave: Minería de Texto; Procesamiento del Lenguaje Natural; Historia Clínica Electrónica; Anamneses.

1. Introduction

The development of information technology (IT) in several fields has offer great deal for increasing the use of electronic medical records in hospitals worldwide. Usually, these EPRs (i.e., anamnesis and evolution of pregnant patients in our study) are written in natural language by health professionals in free text fields. Moreover, they contain clinical information about pregnant women's health, including symptoms, drug usage, clinical exams, and so on (Aramaki et al, 2010).

Retrieving relevant information from text fields always requires careful selection of keywords and drafting of queries (SQL queries). Moreover, it is usually a time-consuming process (Fleuren, 2015; Chu, 2002). In this sense, we need to apply effective techniques able to provide qualitative information from the exploration of large volume of data. Furthermore, these types of techniques also need to obtain unexpected relationships, generating new, useful, and understandable information for different purposes based on the identification of patterns (Hand et al, 2001).

Text Mining is an important technique used to understand the language of written documents, dealing with imprecision, uncertainty, and abbreviation of terms, in addition to the meaning and semantics of words. Methods such as Information Retrieval, Information Extraction and Natural Language Processing (NLP) are examples of Text Mining. NLP involves steps such as morphological analysis, syntactic analysis, and semantic analysis, which provide a robust ground when dealing with this type of documents (Wang et al, 2018; Kreimeyer et al, 2017).

In this sense, we present an approach based on text mining and NLP for extracting, clearing and structuring data obtained from EPRs. We also conduct a comparison analysis of manual data retrieval performed by health professionals (e.g., doctors and nurses) and automatic data retrieval performed by our approach in order to verify whether there is any statistical different between both approaches.

Finally, the EPRs from 2016 e 2020 used in the experiments were obtained from the Januario Cicco maternity hospital located in Northeast of Brazil, and they are composed of personal data (e.g., name, age, marital status, and number of children) and anamnesis (e.g., use of licit and illicit drugs, drug treatments, lab exams, medical examiners comments, patient status, treatment and recommendations). Furthermore, after being extracted data from all EPRS, our system stores it in different relational entities to be used for Data Science and Business Intelligence purposes.

This paper is divided into seven sections and organised as follows. Section Background describes the main concepts related to Text Mining, while Section Related Work presents some important studies in text mining applications. Section Methods describes our proposed method and the data used in the experiments. The experimental methodology is presented in Section Experimental Methodology, while the computational results are illustrated in Section Experimental Results. Finally, Section Conclusion and Future Work presents the main conclusions and some directions for future work.

2. Background

Text mining (TM), also referred to as Text Data Mining (TDM), is an important sub-field of Artificial Intelligence (AI) that uses a variety of methodologies to transform unstructured data (free text) found in documents into normalized, structured data suitable for Machine Learning (ML) algorithms (Hearst, 1999; Antons et al, 2020; Ditzz, 2021). According to Fleuren & Alkema, (2015), the application of TM involves three phases, which can be defined as: information retrieval, named entity recognition and information extraction.

Information retrieval (IR) is considered as being the first step in TM and is performed by querying databases using a set of keywords. In our study IR will be related to the retrieval of medical records (i.e., pregnant women medical records), accordingly Fleuren & Alkema, (2015).

Named entity recognition (NER) is the second step in TM. After performing IR, the set of queried documents can be analysed by search algorithms aiming to identify the occurrence of specific keywords of interest and statements on the relations between the set of keywords. By definition, a named entity is a keyword or a set of keywords that clearly identifies an item or concept. The main idea is to link the keywords that are found in the text to concepts referred in the document (Ratinov, 2009).

After performing the first two phases (i.e., IR and NER), information extraction (IE) algorithms can be applied to detect links between concepts in the text. By doing that TM is able to find valuable knowledge which can be used for further analysis. Nowadays, natural language processing (NLP) based methods are among the most used approaches for extracting knowledge from text (Fleuren, 2015).

Natural Language Processing (NLP) is a methodology that is able to help computers to understand text or speech by simulating the human ability to express ourselves through a natural language such as English, Spanish, or Portuguese (Guida, 1986). NLP systems can analyse a large amount of text data, and also understand concepts within this type of data, being able to extract keywords and relationships. Regarding the huge quantity of unstructured data that is produced every day (e.g., EPRs), this approach has become critical to analysing text-based data efficiently (Kreimeyer et al, 2017).

2.1 Text Pre-processing

In this stage, the extracted data needs to be prepared in order to make it usable, improving its quality and significance for the chosen domain. Some basic terminologies are used in this stage, and can be described as follows:

- Document is a text which describes personal data and anamnesis of a pregnant woman at the maternity hospital;
- Tokens represent words (e.g., "sulfato", "de" e "magnesio") in Portuguese;
- Stop words are basically a set of commonly used words that can be discard without loss of meaning (e.g., "o", "a", "no", "na", "para", and etc., mainly articles and prepositions);
- Tokenization is the process to divide unstructured data into tokens such as words, phrase or keywords;
- Stemming is a process that provides reduction of morphological variations, removing suffixes and prefixes from words but keeping their original forms (e.g., "interessante", "interesse" e "interessado", being all stemmed to "interesse").

2.2 Text Similarity Measures

Text similarity measures have been playing an important role in text mining applications such as information retrieval, text classification, document clustering, text summarization and others (Gomaa, 2013). Words can be similar in two ways lexically and semantically. Lexical similarity can be dealt with String-Based approaches, and Semantic similarity with Corpus-Based and Knowledge-Based approaches. As our work concentrates in Lexical similarity, we will highlight only String-Based similarity measures.

String-based similarity measures are methods used to analyse the level of similarity between string sequences and character composition. In this sense, they can be used in NLP to evaluate the similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison (Oghbaie, 2018). Among the existing measures, we briefly introduce the following ones: Levenshtein distance, cosine similarity and Jaro-Wrinkler.

2.2.1 Levenshtein Distance

The Levenshtein distance computes the distance between two words by counting the minimum number of operations needed to transform one word into the other. An operation can be defined as an insertion, deletion, or substitution of a character, or a transposition of two adjacent characters (Okuda et al, 1976). The Levenshtein distance between two strings *a*,*b* is given by *lev(a,b)*, where:

$$lev_{a,b} = \begin{cases} max(i,j) & if\ min(i,j) = 0 \\ min\begin{cases} lev_{a,b}(\mathrm{i}-1,\mathrm{j})+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases}$$

(1)

2.2.2 Cosine Similarity

The Cosine similarity computes the similarity between two vectors of an inner product space that measures the cosine of the angle between them. The smaller the angle, higher the cosine similarity (Li, 2013). The cosine of two non-zero vectors can be derived by using the Euclidean dot product equation:

$$\Lambda \cdot \mathrm{B} = \|\Lambda\|\|\mathrm{B}\| \cdot \cos\theta$$

(2)

where the cosine similarity, $cos\Theta$, is represented as follows:

$$\cos\theta = \frac{\Lambda \cdot \mathrm{B} \sum_{i=1}^{n} \Lambda_i \, \mathrm{B}_i}{\|\Lambda\|\|\mathrm{B}\|\sqrt{\sum_{i=1}^{n} \Lambda_i^2 \sqrt{\sum_{i=1}^{n} \mathrm{B}_i^2}}}$$

(3)

2.2.3 Jaro-Wrinkler

The Jaro-Wrinkler distance is a string metric for measuring an edit distance (i.e., it is a way of quantifying how dissimilar two words are to one another) between two sequences. The lower this distance for two words is, the more similar the words are. The score is normalized such that 1 means an exact match, and 0 means there is no similarity (Leonardo, 2017). The Jaro-Wrinkler distance can be defined as follows:

$$d = 1 - sim_w$$

(4)

$$sim_w = sim_j + lp(1 - sim_j)$$

(5)

$$sim_j = \begin{cases} 0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m - t}{m}) \end{cases}$$

(6)

3. Related Work

Text mining methods have been applied to clinical free text by both health care practitioners and academic researchers. However, there is no international standard for anamnesis. On top of that, such texts carry medical terminologies from different languages (e.g., English, German and French) and are based on domestic terms. Although our approach performs mining over anamnesis written in Portuguese, we will overview some of the recent results achieved mainly for applying these methods in EPRs.

Most of the works found in the Literature had the main goal of using text mining methods over EPRs for Classification or Clustering purposes. Few of them have proposed the structuring of this type of data in a proper Relational Database by using Named Entity Recognition (NERs), which is also our goal. By doing so, we can use entities and their relationships to structure the entire relational database (i.e., in our case).

3.1 Text Mining Applications in Classification and Clustering Tasks

According to Antons et al. (2020), the majority of the works found in this area assign textual objects like documents or words to predefined categories. In this sense, Downs et al. (2018) developed a natural language processing approach for using in EPRs of adolescents with autism spectrum disorders (ASD) aiming to create a set of rules to classify explicit mentions of suicidality in every document in three categories, such as: positive, negative, or unknown.

Hospital-acquired infections (HAIs) is one of the main issues addressed in hospitals worldwide. Based on this fact, Ehrentraut et al. (2018) proposed the application of Supervised ML techniques to the problem of detecting HAIs by applying NLP for extracting and constructing the database and two well-known learning algorithms, support vector machines (SVMs) and gradient tree boosting (GTB).

Free EPR texts consistently lack of consistent standards. Also, it is often hard to have a balanced number of samples for different types of diseases under analysis. Regarding this problem, Guan et al. (2018) developed a model to generate synthetic text of EPRs called Medical Text Generative Adversarial Network or mtGAN. It is based on the Generative Adversarial Network (GAN) and is trained by an algorithm which uses connectionist reinforcement learning concepts. The synthetic text generated by the mtGAN can be then used in classification tasks.

Differently from Text classification which requires additional information from the researcher (e.g., a pre-classified training data set) as input, text clustering relies only on unsupervised algorithms that group textual content based on similarity. In this sense when dealing with documents (e.g., Electronic Medical Record Text), these documents may need to be represented as document-term matrix and their distances calculated by applying common measures, such as: Euclidean, Cosine, or Manhattan distance (Antons et al., 2020).

According to Wu et al. (2019), the ability to make language patterns visible and comparable is essential to certify whether an NLP model can be adapted to a new task. In this sense, they have proposed a contextualized embedding model to visualize such patterns and provide guidance for reusing NLP models in phenotype-mention identification tasks. Here they applied clustering to create clusters of words or phrases in medical documents, which indicates phenotypes related to people.

3.2 Named Entity Recognition Detection in Text Mining

Named Entity Recognition is an important element in text mining mainly for being able to pre-define single words or multi-word phrases in texts belonging to a specific domain. NER identifies both predefined entities as well as the domain of the entities or the entity types from informal texts (Luo et al., 2015). After single words or multi-word phrases in texts have been recognized, the next step is named entity normalization by assigning suitable identifiers to recognized entities.

Named entity normalization is a challenging process in the Biomedical field mainly because many biological terms have multiple synonyms and term variations, and abbreviations are usually used to refer them (Leaman et al., 2015). In Cho et al. (2017), they proposed an approach for normalizing biological entities, such as disease and plant names, by using word embeddings to represent semantic spaces. For diseases, the National Centre for Biotechnology Information (NCBI) disease corpus and unlabelled data from PubMed abstracts were used to construct word representations. For plants, they manually constructed and unlabelled PubMed abstracts were used to represent word vectors.

As already pointed out earlier, the importance of NER in text mining is considerably high. However, most of the NER processes applied for extracting knowledge from EPRs are carried out in English, making reuse difficult. Based on this fact, Grechishcheva et al. (2019) have proposed a work focused on applying automatic pre-processing methods on EPRs data written in Russian using specially developed domain knowledge base (library).

According to them, Concepts and terms from library are used for deciphering abbreviations and acronyms; this step increases semantic accuracy. Also, the library contains marked tokens, which can be removed as they are not important in the context of clinics and treatment processes; this step reduces detentions of text vectors. The computational experiment was conducted on real data from 41,025 EPRs of patients, who were diagnosed with stroke and treated at the Almazov National Medical Research Centre of Saint Petersburg from 2010 to 2019.

3.3 Discussion

As already pointed out, there is no international standard for extracting text from anamnesis which is an important part of an EPR that describes relevant information of a patient, such as: use of licit and illicit drugs, drug treatments, lab exams, medical examiners comments, patient status, treatment, and recommendations. On top of that, such texts carry medical terminologies from different languages (e.g., English, German and French) and are based on domestic terms, making reuse difficult.

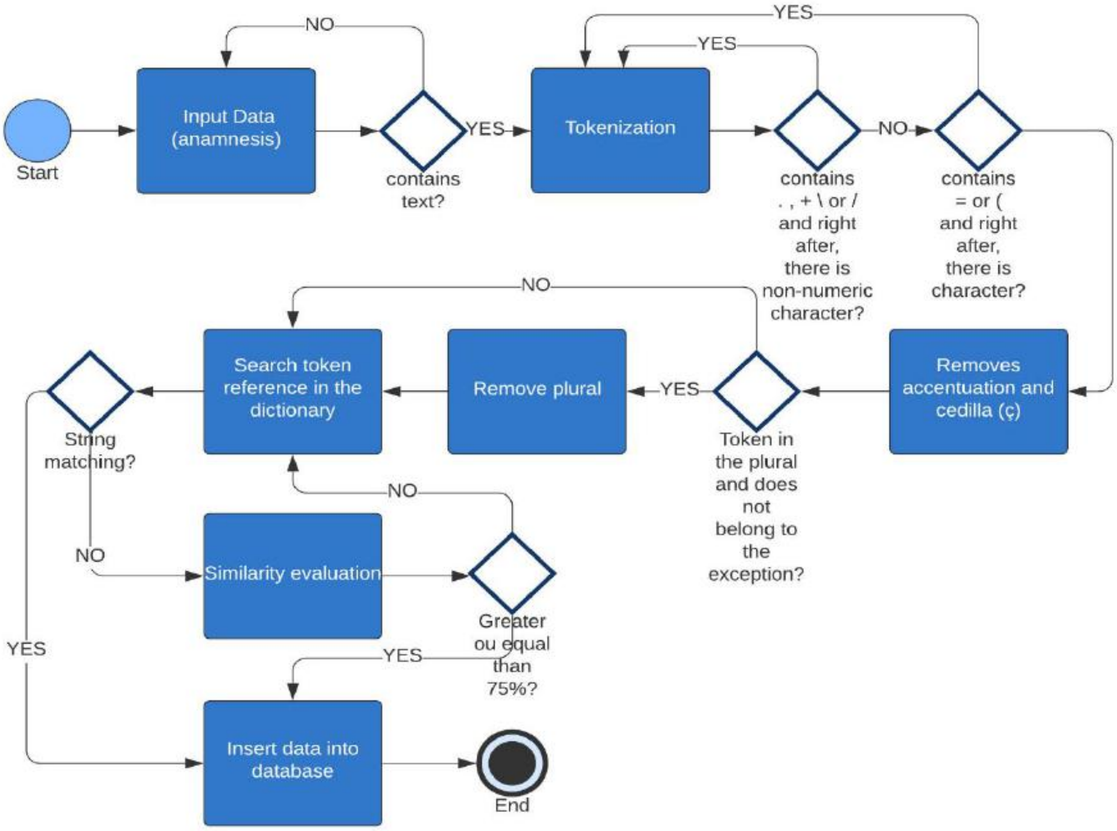
Based on this important fact, and also considering our needs, we propose an approach based on text mining and NLP for extracting, clearing and structuring data obtained from EPRs. Our NER process aims to identifies, extracts and structures terms related to all relevant information of a patient, storing them into relational database for further statistical analysis.

Moreover, our approach differs from all related work in two aspects: (1) as the anamnesis are written in Portuguese, a new dictionary has been defined for covering all the terms related to the patients; (2) as most of the works use NLP methods to generate datasets for classification purposes, our approach extracts terms and stores them for further analysis.

4. Methods

The clinical data (anamnesis) was provided by the Januario Cicco maternity hospital which belongs to the Federal University of Rio Grande do Norte located in northeast of Brazil. This maternity hospital receives medical students in their last year, as well as post-graduate students in the areas of obstetrics and gynecology.

Figure 1. Flow of our method created to perform the structuring of the anamnesis.


Start
Input Data (anamnesis)
NO
contains text?
YES
Tokenization
YES
YES
contains . , + \ or / and right after, there is non-numeric character?
NO
contains = or (and right after, there is character?
Removes accentuation and cedilla (ç)
Token in the plural and does not belong to the exception?
YES
Remove plural
NO
Search token reference in the dictionary
String matching?
NO
Similarity evaluation
Greater ou equal than 75%?
NO
YES
YES
Insert data into database
End

Source: authors.

The data used in our study includes 4,000 anamneses of women in both areas of obstetrics and gynecology from 2016 e 2020. The clinical data is composed of the following information: name, age, marital status, number of children, use of licit and illicit drugs, drug treatments, lab exams, medical examiners comments, patient status, treatment, and recommendations. In order to extract and structure all the relevant information from the anamneses, we created an automatic method implemented in Java which is is presented in Figure 1.

As it can be seen in our method's flow, non-structured data (i.e., text from PDF files written in Portuguese) is loaded from the Januario Cicco maternity hospital's medical information system to the starting point of the algorithm. Firstly, the algorithm checks whether the files (PDF's) are readable (i.e., contain text). In case of non-readable file, this file is discarded, and the next file is loaded. The next step of the flow is related to pre-processing of text.

4.1 Data Pre-processing

After checking the files, the algorithm performs the tokenization step which generates tokens based on plain text. In order not to disconnect the relationship between composed words and their semantic content, the sequence of the tokens was maintained. Thus, we analyse tokens individually, and also the combination of its predecessors' tokens for better understanding. Based on this fact, we did not apply any stop-word removal, avoiding losing part of the information.

Secondly, we implemented two steps to split words that remained joined together for any reason. Step A searches for digits after finding anyone of the following symbols: "+", ",", ".", "\" and "/", aiming to split words written without spacing (e.g., "AU = 36cm", "ACF = 130bpm"). Step B splits any character after finding: ":", "=" and "(".

Finally, a morphological analysis was applied, aiming to remove diacritics marks (e.g., "ão", "í", "ç") in the Portuguese words. Although it is natural to use Stemming algorithm in this process, we decided not to use it but only transform words from plural to singular form, making exception for few tokens, such as: "aids", "has", "syphilis", "herpes" and "diabetes". Table 1 presents four rules used to transform words from plural to singular form accordingly to Portuguese grammar.

Table 1. Description of degree courses, courses, activities, and resources extracted from Moodle.

Rule/Solution	Description
Rule 1	Words ending with "ões", "ãos" or "ães"
Solution 1	Transform them in "ao"
Rule 2	Words ending with "es" and preceded by "l", "r", "s", or "z"
Solution 2	Remove "es"
Rule 3	Words ending with "is" and preceded by vowel
Solution 3	Remove "is" and add "l" at the end
Rule 4	Words ending with "s"
Solution 4	Remove "s"

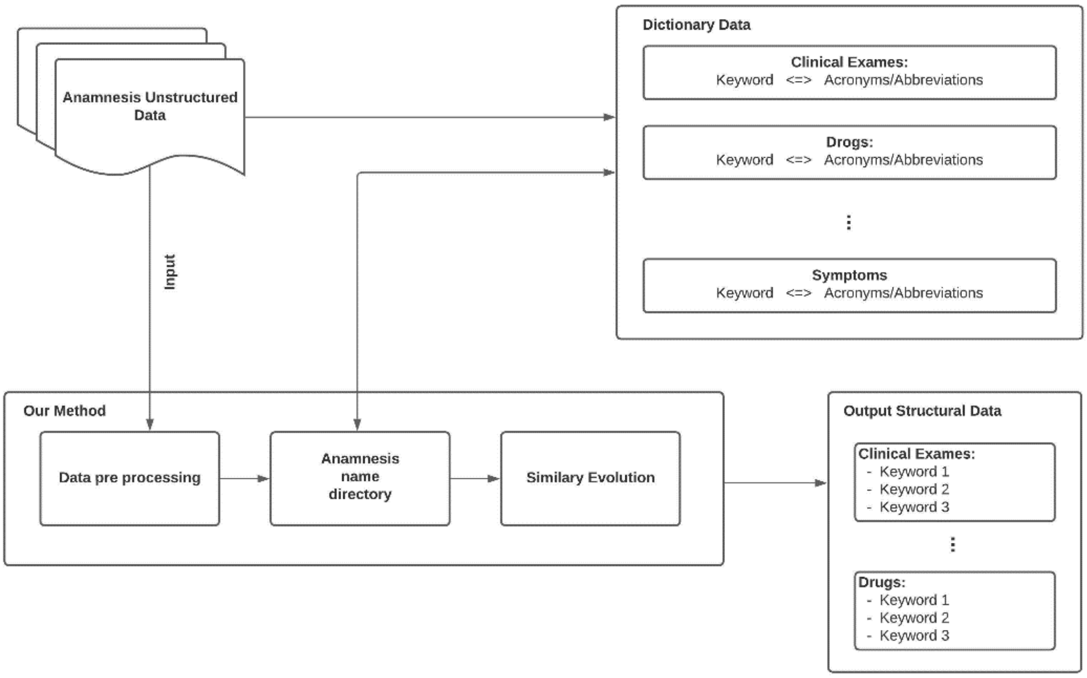
Source: authors.

4.2 Anamnesis Name Dictionary

Aiming to define and create an anamnesis dictionary, a subset of one thousand (1000) anamneses were evaluated by a team composed by 10 health professionals: two doctors, six nurses and two pharmacists. All relevant data identified by the team was classified into 12 categories (see right side of Figure 2), being: presence of symptom, symptom denial, medication in use, medication administered, prescribed medication, applied conduct, co-morbidity / disease found, use of licit and illicit drugs, referral between institutions / cities, medical exams, allergy, and general characteristics such as blood type and Rh factor, multiple delivery, and tubal ligation request.

The aforementioned categorization was based on a famous obstetrics book \cite{Montenegro2014} used as a foundation book on obstetrics that defines the evaluation process in patient care, as well as base for Resolution No. 2056/2013 of the Federal Council of Medicine of Brazil.

Figure 2. The functioning of the proposed automatic method.



Source: authors.

Secondly, we normalized all relevant information, specifying the relation between keywords and their acronyms and abbreviations, as well as single and compound words. In this sense, when a token is identified synonymous with a keyword, the relationship is set. The dictionary construction process was carried out manually, analysing information from 1000 anamneses (i.e., 25\% of the original data).

4.3 Similarity Evaluation

In case of not reaching string matching, a token had its similarity evaluated against all the keywords defined in the Anamnesis dictionary. Regarding the string-based similarity measure, which was discussed in the Background section, we analysed the performance of all three, being Levenshtein distance, cosine similarity and Jaro-Wrinkler. Based on the results, we decided to apply the Levenshtein distance setting a threshold greater or equal than 75%.

5. Experimental Methodology

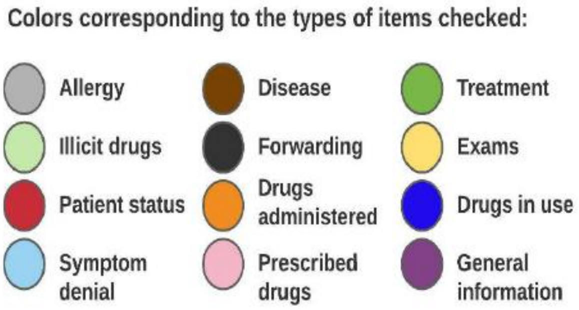
In order to evaluate our proposed method for extracting and structuring data from anamneses, we selected 3,000 thousand (i.e., 75\%) anamneses from the original data. Moreover, none of these anamneses were used in the construction phase of the Anamnesis name dictionary. By doing so, we believe that no bias was introduced in this analysis, since there was no intersection between different subsets.

The methodological experiment was carried out through the invitation of 30 medical professionals from the maternity hospital, being: 10 doctors and 20 nurses. Each one of them received 100 anamneses randomly selected, and manually classified all relevant data according to the 12 categories specified in the Anamnesis name dictionary.

5.1 Evaluation Process

All 30 medical professionals received different colours of markers to identify the categories, as shown in Figure 3. In this sense, we were able to conduct a comparison analysis between the manual classification performed by the professionals and the automatic classification performed by our method. The main purpose of this analysis was to investigate whether our method was able to reach a good and competitive accuracy compared to human skills.

Figure 3. Legend with the evaluated categories.



Source: authors.

Three different groups were formed in order to evaluate the results, being group A composed by doctors; group B composed by nurses; and group C composed by our method. The same team (i.e., 10 health professionals) of experts in both areas of obstetrics and gynecology that specified and created the Anamnesis name dictionary checked the results, assigning one point to a correct information marked accordingly to its category.

The average of the correct markers was computed by dividing the total number of correct markers for each group and the maximum number of tokens present in each anamnesis. In this sense, it was generated three averages of correct markers, one for each group.

5.2 Comparative Analysis

In order to conduct a statistical analysis of the results, the Kruskal-Wallis (Kruskal, 1952) test and Bonferroni post-hoc test were used to determine whether or not there is a statistically significant difference between all three groups. Both tests are applied at the conventional significance level of 5%.

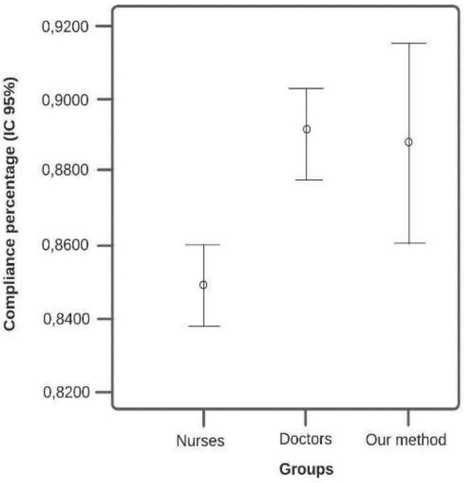
6. Experimental Results

Figure 4 presents the statistical results for all three groups (doctors, nurses, and our method) when analysing the tokens of 3,000 anamneses and their corresponding categories. As it was expected, the group A composed by 10 doctors obtained the best results (89,13 ± 8,85). Group C represented by our method obtained the second-best results (88.91 ± 7.77), coming third group B composed by 20 nurses (84.87 ± 9.18).

Kruskal-Wallis test produced the *p-value* = 0.0208, therefore the difference between the average of correct markers of the three groups is statistically significant. However, the pairwise comparisons using the Bonferroni post-hoc test produced only one statistically significant result: group A obtained a significantly better average of correct markers than group C (*p-value* = 0.0420). That is, there is no significant difference between the average of correct markers of other pairs of groups.

Moreover, although there was no significant difference between groups A and C (*p-value* = 1.000), and between groups C and B (*p-value* = 0.0600), group B (our method) presented similar results in comparison to group A (doctors). On top of that, group C presented the smallest standard deviation value of all three groups.

Figure 4. Statistical results by group.



Source: authors.

6.1 Discussion of the Results

Based on the results, it appears that the group of nurses was more focused on checking medications, allergies and clinical exams of the patients, contributing to a lower performance compared to the two other groups. In fact, during the care routine nurses are responsible for managing patient care, which emphasize the type of behaviour noticed throughout the evaluation of the anamneses. The doctors, on the other hand, are more concern with the patient's history, aiming to identify what is causing the problem. In this way, they usually check the general information, analysing each point of an anamnesis.

Regarding the results obtained by our method, it is important to emphasize that it had a poor result when identifying clinical exams. Therefore, it necessary to expand the Anamnesis name dictionary with more keywords related to this category.

7. Conclusion and Future Work

This work proposed a method which applies Text Mining and Natural Language Processing to Electronic Patient Records (EPRs) for extracting and transforming texts into structured data. In order to achieve our goal, we 4,000 anamneses (i.e., clinical data) of women in both areas of obstetrics and gynecology from 2016 e 2020 provided by the Januario Cicco maternity hospital which belongs to the Federal University of Rio Grande do Norte located in northeast of Brazil.

Based on the results presented and discussed in Section Experimental Results, our method obtained a good and competitive accuracy compared to human skills. In fact, considering group A (doctors), group B (nurses) and group C (our method), group A obtained the best results (89,13%) as expected, group C came second (88,91%) and group B third (84,87%). Moreover, all the relevant information retrieved from the 4,000 anamneses will be now available for further analysis (e.g., BI dashboards).

As a direction for future work, it is necessary to expand the Anamnesis name dictionary with more keywords related to clinical exams. Moreover, as to the best of our knowledge there is no generic anamnesis dictionary in public and private health care service in Portuguese. In this sense, we intend to invite medical professionals from other areas to contribute for creating a generic dictionary applicable for many areas in health care service.

References

Antons, D., Grünwald, E., Cichy, P. & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, 50(3), 329-351.

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., Waki, K. & Ohe, K. (2010). Extraction of Adverse Drug Effects from Clinical Records. In *Proceedings of the 13th World Congress on Medical (MEDINFO 2010*) (pp. 739-743). IOS Press.

Cho, H., Choi, W. & Lee, H. (2017). A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*, 18(451), 1-12.

Chu, S. (2002). Information retrieval and health/clinical management. *Yearbook of medical informatics*, 1, 271–275.

Ditzz, A J. M.; Gomes, G. R. R. (2021). Tool for developing self-assessment tests for distance learning environments. *Research, Society and Development*, 10(8), p. e32710817424, 2021. DOI: 10.33448/rsd-v10i8.17424.

Downs, J., Velupillai, S., George, G., Holden, R., Kikoler, M., Dean, H., Fernandes, A. & Dutta, R. (2018). Detection of suicidality in adolescents with autism spectrum disorders: Developing a natural language processing approach for use in electronic health records. *Journal of the American Medical Informatics Association*, 641-649.

Ehrentraut, C., Ekholm, M., Tanushi, H., Tiedemann, J. & Dalianis, H. (2018). Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. *Health Informatics Journal*, 24(1), 24–42.

Fleuren, W. W. M. & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, 74, 97–106.

Gomaa, W. & Fahmy, A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68, 13–18.

Grechishcheva, S., Efimov, E. & Metsker, O. (2019). Risk markers identification in EHR using natural language processing: hemorrhagic and ischemic stroke cases. *Procedia Computer Science*, 156, 142–149.

Guan, J., Li, R., Yu, S., & Zhang, X. (2018). Generation of synthetic electronic medical record text, In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 374–380.

Guida, G. & Mauri, G. (1986). Evaluation of natural language processing systems: Issues and approaches. *Proceedings of the IEEE*, 74(7), 1026–1035.

Hand, D.J., Smyth, P. & Mannila, H. (2001). Principles of Data Mining. *MIT Press*, Cambridge, MA, USA.

Hearst, A. (1999). Untangling text data mining, In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99*, 3–10, USA: Association for Computational Linguistics.

Leaman, R., Khare, R. & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57, 28–37.

Leonardo, B. & Hansun, S. (2017). Text documents plagiarism detection using rabin-karp and jaro-winkler distance algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 5(2), 462–471.

Li, B. & Han, L. (2013). Distance weighted cosine similarity measure for text classification. *Intelligent Data Engineering and Automated Learning*, 8206, 611–618.

Luo, G., Huang, X., Lin, C.Y.& Nie, Z. (2015). Joint entity recognition and disambiguation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 879–888, Lisbon, Portugal: Association for Computational Linguistics.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M. & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: Asystematic review. *Journal of Biomedical Informatics*, 73, 14–29.

Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.

Montenegro, C. A. B. & Rezende, J.F. (2014). *Fundamental Obstetrics*, 13th edition, Gen.

Oghbaie, M. & Mohammadi, Z. M. (2018). Pairwise document similarity measure based on present term set. *Journal Big Data*, 5(52), 1–23.

Okuda, T., Tanaka, E. & Kasai, T. (1976). A method for the correction of garbled words based on the Levenshtein metric. IEEE Transactions on Computers, C-25(2), 172–178.

Okuda, T., Tanaka, E. & Kasai, T. (1976). A method for the correction of garbled words based on the Levenshtein metric. *IEEE Transactions on Computers*, C-25(2), 172–178.

Ratinov, L. & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, 147–155, Colorado: Association for Computational Linguistics.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S. & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77, 34–49.

Wu, H., Hodgson, K., Dyson, S., Morley, K. I., Ibrahim, Z. M., Iqbal, E., Stewart, R., Dobson, Richard, J.B., & Sudlow, C. (2019). Efficient reuse of natural language processing models for phenotype-mention identification in free-text electronic medical records: A phenotype embedding approach. *JMIR Med Inform*, 7(4), e14782.