Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 684–689

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems (HiPIS 2019)
November 4-7, 2019, Coimbra, Portugal

Automatically detect diagnostic patterns based on clinical notes through Text Mining

João Ribeiro[a], Júlio Duarte[b], Filipe Portela[b*], Manuel F. Santos[b]

[a] *Department of Information Systems, University of Minho, Campus de Azurém, Guimarães 4800-058, Portugal*
b Centro Algoritmi, University of Minho, Campus de Azurém, Guimarães 4800-058, Portugal

Abstract

The importance of standardized treatment for patients is huge because it can reduce waiting times, costs in hospitals and make treatment more effective for patients. According to these patterns, the creation of a tool that can make the admission and interpretation of free text will become an important step in the medical field. For the analysis of the unstructured text, the "RapidMiner" tool was used. Following the text analysis, the word frequency technique will be used in the reports and the respective word counts, as well as the cluster analysis that allows the creation of combinations of words. For the modeling we used several Text Mining techniques focused on the main algorithms, since these are properly scientifically proven and that, normally, they are able to obtain better results.

Keywords: Text Mining; Text Analysis;

1. Introduction

In the scope of this study, the following research question can be identified throughout the article: Will it be possible, using Text Mining techniques applied to a set of reports of Computed Tomography (CT) tests, to induce models that classify the associated diagnosis?

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: cfp@dsi.uminho.pt

10.1016/j.procs.2019.11.027

In this study we intend to find logical patterns for patients' diagnoses based on the clinical notes obtained as well as assign a major role to Text Mining and Natural Language Processing (NLP) in the health area.

The primary goal is to interpret clinical information that will automatically interpret what is written about a patient and help doctors make a quick and effective decision. In this regard, it will be necessary to create a clinical dictionary with the words used by physicians on a daily basis so that it is subsequently possible to find important patterns in patients' clinical notes. The models will be produced and tested through the induction of Text Mining models, using real data provided by a Portuguese hospital. In order to achieve this objective, it is necessary to create small objectives that complement this final objective, as such will be necessary: a) Translation of clinical notes based on natural language of data useful for analysis; b) Creation of a system capable of automatically detecting clinical patterns; c) A tool capable of analyzing and interpreting clinical notes; d) New algorithms for interpreting clinical information; e) A clinical dictionary; f)A new knowledge in the area of Information Systems applied to health. The fulfillment of these objectives becomes central to the achievement of the main objective.

2. Background

2.1. Text Mining

Text Mining seeks to extract useful information from data sources by identifying and exploring patterns. Text Mining, also known as "Text Data Mining" (TDM) or knowledge discovery of "Text Data Base", generally refers to the process of extracting patterns or non-trivial knowledge of documents of unstructured text and can be seen as an extension of Data Mining. The Text Mining system receives as inputs a set of raw documents and generates various types of outputs, such as patterns, connection maps and trends [1] [4] [6].

There are a number of possible Text Mining techniques to be used in the different application areas, so Chauhan Shrihari and Amish Desai then define the techniques that, in their perspective, are the most used [2]: 1. Extraction of Information; 2. Clustering; 3.Summarization; 4.Visualization; and 5. Categorization.

The patient's medical records contain a wealth of information that may be necessary for the conduct of clinical research. Clinical notes written by doctors are written in free text format. Therefore, it is necessary to use information extraction techniques that will later allow us to use a reliable and efficient method to extract structured information from the Data Mining that, in the end, are directed to achieve reduced profits with the research efforts.

2.2. Natural Language Processing

Natural Language Processing (NLP) is a research and application area that explores how computers can be used in order to understand and even be able to manipulate natural language text.

The natural language processing is a set of computational techniques motivated by interest of analysis and representation of texts, with the objective of obtaining human-like language processing for a variety of tasks or applications [9]. The different levels of natural language processing are [9]: 1. Phonology that deals with pronunciation; 2. Morphology that deals with the smallest parts of the words, which direct a meaning and yet, suffixes and prefixes; 3. Lexical that deals with the lexical meaning of words and parts of speech analysis; 4.Syntactic that deals with grammar and sentence structure; 5. Semantic that deals with the meaning of words and phrases; 6. Discourse that deals with the structure of different types of text using document structures; 7.Pragmatic that deals with the knowledge that arises from the outside world, that is, outside the contents of the document.

A system of Comprehension of Natural Language is capable of [9]: a) Explain a text; b) Translate the text into another language; c) Answer questions about the content of the text; and d)Make deductions about the text.

The purpose of the NLP system is to represent the true meaning and intent of the user query [9].

2.3. Clinical Notes

Clinical notes are an essential component for the treatment of a patient, whereby the information must be accurate, objective and necessary as well as consistent with the objective in question.[21] The clinical notes are presented as a record containing information on health status, personal history, current medication, contacts with

health services, as well as examinations, therapies and surgeries already performed. The Electronic Health Record (EHR) can be defined as a clinical information system that supports the needs of health professionals in all clinical hospital departments and their functional areas. The use of standard terminologies for registration and clinical research allows the creation of scenarios from the EHR repository, favoring the management and support of intelligent decision making. Thus, EHR is also a starting point for the implementation of intelligent systems in hospital units [25].

3. Methods and Tools

For the analysis of the unstructured text, the RapidMiner tool was fully accepted and recognized in the community, and was characterized as quite complete, feasible and scientifically proven.

Regarding the elaboration of the entire project, two methodologies are essentially used, the first one to assist the investigation (Design Science Research (DSR) [12]) and the second (Cross Industry Standard Process for Data Mining (CRISP-DM) [13]) to guide the development project in Text Mining. Thus, for a correct use of the methodologies, they will be duly respected and fulfilled in all phases of the project.

4. Business Study

We intend to use the clinical notes in order to define predictive patterns and thus we can reduce waiting time and the possibility of failure. To address these two drawbacks, several predictive models were designed to predict the diagnosis of each patient based on the clinical notes provided. Following this objective, it was proposed to use the keywords provided by the Portuguese hospital unit in order to understand the utility and importance in the forecast. Knowledge of Text Mining, Data Mining, ability to interpret data as well as business interpretation and some technical knowledge to use the necessary tools is required. To allow a huge breakthrough in the field of medicine as it will provide a very high reduction in waiting time as well as the reduction of human failure, since they are two fundamental factors in this area. It will now be feasible to use all the unstructured text produced by physicians in order to discover and fit the existing patterns between words, which has many benefits in the future in medicine. This is due to the fact that all freely written text is never used for the preparation of this type of analysis.

5. Data Study

5.1. Data Preparation

The data do not present structured text in relation to those who underwent CT examinations as well as to their diagnosis. The data provided is in a dataset called "export_diag_relats" containing 7138 records. In addition to the dataset described above, the dataset of the keywords that were later used to analyze and predict the diagnoses, the "export_diag_reg_exp" containing 26 registers was also provided. In order to optimize the whole analysis, the need to use the Oversampling technique has arisen, that is, to replicate the diagnoses that have the smallest registers, as many times as necessary, so that all of them are balanced, thus allowing greater efficiency, better results. That is, with the execution of Oversampling the mentioned TOP 5 contains the following number of records:

1. The diagnosis "434 - Occlusion of the cerebral arteries" contains 225 associated reports;
2. The diagnosis "852 - Subarachnoid hemorrhage" contains 214 reports;
3. Diagnostic records "432 - Other intracranial hemorrhages" replicated four times (51 registers x 4 = 204);
4. The diagnostic records "435 - Brain ischemia" were replicated four times (48 registers x 4 = 192 registers).
5. Diagnostic records "851 - Laceration and contusion" replicated five times (39 records x 5 = 195 records).

5.2. Keyword List

The table below lists the keywords used in the study.

Table 1. Keyword list.

Diagnosis	Keywords
434-Oclusão das artérias cerebrais	Oclusão; Artérias; Cerebrais;
852-Hemorragia subaracnoídea	Hemorragia; Subdural; Extradural;
432-Outras hemorragias intracranianas	AVC; Vascular; Cerebral;
435-Isquemia cerebral	Isquemia; Cerebral;
851-Laceração e contusão	Laceração; Contusão; Cerebral;

6. Results

6.1. Frequency of words

In the table 2. it is possible to see an excerpt from the list of keywords ordered by their frequency in the reports.

Table 2. Frequency of words.

Word	Attribute	Total
Cerebral	Cerebral	715
Subdural	Subdural	397
Artérias	Artérias	241
Contusão	Contusão	174
Hemorragia	Hemorragia	151
Vascular	Vascular	140
Cerebrais	Cerebrais	121
Hemorragias	Hemorragias	65
Intracranianas	Intracranianas	46
Oclusão	Oclusão	42

6.2. Clusters creation

The five clusters you create identify the most commonly used expressions in the reports provided that consist of the following keywords:

- Cluster 0: “Artérias, Cerebral”
- Cluster 1: “Subdural, Cerebral e Hemorragia”
- Cluster 2: “AVC, Vascular, Cerebral, Hemorragia, Intracraniana, Laceração, Isquemia”
- Cluster 3: “Cerebral, Oclusão”
- Cluster 4: “Contusão, Cerebral, Hemorragia, subdural”

6.3. Modeling

Like can be verified that the result set is unsatisfactory, since several faults have been detected in the results set. In Scenario 1, the "Ratio", only the "K-Nearest Neighbor" algorithm obtained less unsatisfactory results, the only one to show values close to 75%. In Scenario 2, the "Binary", no results are satisfactory since the hit rate found for all algorithms is approximately 50%. In scenario 3, the "Count", the results obtained are presented as unsatisfactory since for the four algorithms used in the creation of the models, there is again the possibility of error of approximately 50% and in some of them this value may still be higher. As in the last two previous scenarios, in scenario 4, the "WordRatio", the results of the models remain unsatisfactory with a probability of error of approximately 50%. Below is the matrix with all the results of the respective combinations.

Table 3. Results

Techniques	Algorithm	Accuracy
Ratio	Decision Tree	54,92+/- 3,14%
	K-NN	74,13+/- 4,61%
	Neural Networks	48,64+/- 3,32%
	LibSVM	47,18+/- 6,25%
Binary	Decision Tree	49,80+/- 3,86%
	K-NN	41,50+/- 2,86%
	Neural Networks	51,25+/- 4,41%
	LibSVM	49,61+/- 3,89%
Count	Decision Tree	46,24+/- 3,84%
	K-NN	51,64+/- 3,49%
	Neural Networks	51,25+/- 6,23%
	LibSVM	31,56+/- 2,03%
WordRatio	Decision Tree	45,66+/- 4,03%
	K-NN	46,33+/- 2,51%
	Neural Networks	51,54+/- 5,22
	LibSVM	46,91+/- 4,33%

7. Discussion

In short, only one model was able to achieve results that could still be used by any forecasting system, however, in order to obtain a satisfactory result, it would be necessary to make an investment in model optimization. However, all the other models present opposite results, which means that the keywords indicated by the Portuguese hospital unit to predict a particular diagnosis are not the most appropriate, but it is possible that there are some words with a high degree of importance implying better results in the predictions of the diagnoses.

It is concluded that the "K-Nearest Neighbor" algorithm of the first scenario is the only one that can be considered acceptable although not satisfactory, since in medicine approximately 25% error can be fatal, since the object of study are human beings, that is, it is completely crucial to cancel the error with the maximum certainty, being that only an efficient system could guarantee these purposes.

In order to answer the question, "will it be possible, by using Text Mining techniques applied to a set of computed tomography (CT) scan reports, to induce models that efficiently classify the associated diagnosis?" several forecasting models were created using the RapidMiner tool. The results show that only the "K-Nearest Neighbor" algorithm of the first scenario is the only one that can be considered acceptable although it is not satisfactory, since it is decisive to cancel the error with the maximum certainty in the area of medicine.

8. Conclusion

In order for the main purpose to be fulfilled, the focus would be on the creation of a diagnostic forecasting model and, therefore, the majority need to fulfill the small objectives initially proposed. The translation of clinical notes based on natural language of the data useful for the analysis, through the use of the RapidMiner tool, allowed all the treatment of the unstructured text to obtain useful information for the analysis. Thus, with the creation of text analysis models based on the use of clusters, it was feasible to detect the existing clinical patterns in an automatic way, since they are to interpret the clinical information in order to demonstrate the results referring to the same. In a possible future work, this project provides a continuation of the main objective of identifying which keywords can predict the selected diagnoses, thus guaranteeing very satisfactory prediction models, however, leads to different conclusions. However, it is not fully guaranteed that such keywords exist because reports are free-text, so there is no mandatory logical pattern to follow.

Currently, forecasting is extremely useful as it can reduce downtime and the drastic possibility of errors or failures that may exist, making it an advantage for the organization as well as for society in general.

In short, it will be very beneficial and rewarding to continue all medical research, as there are many records that are not structured but contain useful information on the provision of aid, for example to support decision making.

Acknowledges

This work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019and Deus ex Machina (DEM): Symbiotic technology for societal efficiency gains - NORTE-01-0145-FEDER-000026.

References

[1].Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Imagine, 34, 410. https://doi.org/10.1179/1465312512Z.00000000017

[2]. Shrihari, C., & Desai, A. (2015). A Review on Knowledge Discovery using Text Classification Techniques in Text Mining. International Journal of Computer Applications, 111(6), 975–8887. Retrieved from http://research.ijcaonline.org/volume111/number6/pxc3900784.pdf

[3]. Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. Information Processing & Management, 43(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

[4]. Tan, A. H. (1999). Text mining: The state of the art and the challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, 65–70. https://doi.org/10.1.1.38.7672

[5]. Hearst, M. A. (1999). Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics - (pp. 3–10). https://doi.org/10.3115/1034678.1034679

[6]. Truyens, M., & Van Eecke, P. (2014). Legal aspects of text mining. Computer Law & Security Review, 30(2), 153–170. https://doi.org/10.1016/j.clsr.2014.01.009

[7]. Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence. https://doi.org/10.4304/jetwi.1.1.60-76

[8]. Piedra, D., Ferrer, A., & Gea, J. (2014). Text Mining and Medicine: Usefulness in Respiratory Diseases. Archivos de Bronconeumología (English Edition), 50(3), 113–119. https://doi.org/10.1016/j.arbr.2014.02.008

[9]. Liddy, E. D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science (Vol. 37, pp. 51–89). https://doi.org/10.1017/S0267190500001446

[10]. Gobinda G. Chowdhury. (2003). Natural Language Processing. Annual review of information science and technology (Vol. 37). https://doi.org/10.1017/S0267190500001446

[11]. Kao, A., & Poteet, S. R. (2007). Natural language processing and text mining. Natural Language Processing and Text Mining. https://doi.org/10.1007/978-1-84628-754-1

[12]. Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. J. Manage. Inf. Syst., 24(January), 45–77. https://doi.org/10.2753/MIS0742-1222240302

[13]. Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data minning guide. CRISP-DM Consortium, 76.

[14]. Jha, A. K., DesRoches, C. M., Campbell, E. G., Donelan, K., Rao, S. R., Ferris, T. G., … Blumenthal, D. (2009). Use of electronic health records in U.S. hospitals. New England Journal of Medicine, 360(16), 1628–1638. https://doi.org/NEJMsa0900592 [pii]\r10.1056/NEJMsa0900592

[15]. Zhou, X., Han, H., Chankai, I., Prestrud, A., & Brooks, A. (2006). Approaches to Text Mining for Clinical Medical Records. Proceedings of the 2006 ACM Symposium on Applied Computing, 235–239. https://doi.org/10.1145/1141277.1141330

[16]. Peter Suber. (2009). Open Access Overview. Exploring Open Access: A Practice Journal. https://doi.org/10.1109/ASPDAC.2006.1594722

[17]. Garde, S., Knaup, P., Hovenga, E. J. S., & Heard, S. (2007). Towards semantic interoperability for electronic health records: Domain knowledge governance for openEHR archetypes. Methods of Information in Medicine, 46(3), 332–343. https://doi.org/10.1160/ME5001

[18]. Freitas, F., Schulz, S., & Moraes, E. (2009). Pesquisa de terminologias e ontologias atuais em biologia e medicina. Reciis, 3(1), 8–20. https://doi.org/10.3395/reciis.v3i1.239pt

[19]. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a Web of open data. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 4825 LNCS, pp. 722–735). https://doi.org/10.1007/978-3-540-76298-0_52

[20]. Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2), 159–165. https://doi.org/10.1147/rd.22.0159

[21]. Lyndon K. (2015). Best Practice Clinical Note Taking. Exercise & Sports Science Australia. https://www.essa.org.au/wp-content/uploads/2015/04/BEST-PRACTICE-CLINICAL-NOTE-TAKING.pdf

[22]. Garde, S., Knaup, P., Hovenga, E. J. S., & Heard, S. (2007). Towards semantic interoperability for electronic health records: Domain knowledge governance for openEHR archetypes. Methods of Information in Medicine, 46(3), 332–343. https://doi.org/10.1160/ME5001

[23]. Garde, S., Hovenga, E., Buck, J., & Knaup, P. (2007). Expressing clinical data sets with openEHR archetypes: A solid basis for ubiquitous computing. International Journal of Medical Informatics. https://doi.org/10.1016/j.ijmedinf.2007.02.004

[24]. Dolin, R. H., Alschuler, L., Beebe, C., Biron, P. V., Boyer, S. L., Essin, D., … Mattison, J. E. (2001). The HL7 clinical document architecture. Journal of the American Medical Informatics Association. https://doi.org/10.1136/jamia.2001.0080552

[25]. A. Abelha, J. Machado, J. Neves (2008), O processo clínico eletrónico. http://hdl.handle.net/1822/19005