2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Preliminary Text Analysis from Medical Records for TB Diagnosis Support

Andrés Felipe Romero Gómez, Alvaro D. Orjuela-Cañón, *Member, SMIEEE*, Andrés L. Jutinico, Carlos Awad, Erika Vergara, and Angélica Palencia

***Abstract*— Tuberculosis is an infectious disease that is spread through the air from one person to another and is one of the top ten causes of death in the world according to the World Health Organization. From biomedical engineering, decision support systems based on artificial intelligence have shown advantages for healthcare personnel in tasks such as diagnosis and screening. A specific area of the artificial intelligence is the natural language processing, however, most of these approaches are based on available data. This paper shows the construction of a dataset based on medical records of subjects suspected of tuberculosis. In addition, an initial exploration of the contents of the constructed dataset and how this approach can be followed by a natural language processing to support tuberculosis diagnosis in data demanding scenarios are presented.**

***Clinical Relevance*— In some developing countries as Colombia, it is difficult to develop systems based on artificial intelligence due to the availability of data. This proposal holds a strategy to build a dataset to train machine learning models, and to obtain support diagnosis tools, employing natural language from the medical scenario from text written by health professionals in the medical record. In this way, trained models based on this information available can be employed in places where medical infrastructure is precarious.**

I. Introduction

Tuberculosis (TB) is an infectious disease caused by the *Mycobacterium tuberculosis*. This bacterium mostly attacks the lungs, but can also affect other parts of the body, and it can be easily spread between people through the air, especially in areas with high population density values and low socioeconomic conditions [1]. Additionally, TB has been recognized as a universal emergency by the World Health Organization (WHO) due to its worldwide impact, and it is among the top 10 leading causes of death by a single infectious agent, thus, ending the TB pandemic by 2030 is one of the health-related targets of the Sustainable Development Goals (SDGs) [2].

For developing countries, the situation is especially difficult. Detection of TB is challenging due to the limitations of the medical infrastructure [1] [3]. Specialized laboratories are demanded to determine if TB suspected people holds the disease. In addition, health professionals to cover regions faraway from big cities represent a problem in places with basic structure for public health. In this way, Colombia as a Latin American country, holds a TB incidence that fluctuates during the last 10 years with a certain tendency to increase. For 2019, it had an incidence of 26.9 per 100 thousand inhabitants. This increasing behavior can be explained by the strengthening of surveillance and monitoring actions of the disease that have been carried out in the country [4].

From biomedical engineering, different strategies in the health area have been increased with new proposals to solve traditional problems. In this way, applications based on artificial intelligence (AI) techniques are being developed, highlighting the so-called decision support systems (DSS). These techniques have been shown to be useful as support tools in tasks for the diagnosis and prognosis of diseases, providing an extra-help to health professionals, contributing more and new sights to treat the problems [5] [6] [7].

For the specific case of TB diagnosis support, AI has been employed in different scenarios with the use of artificial neural networks in demanding scenarios [8], proposals based on images [9], and other applications [10], [11]. Furthermore, recent applications of the AI are related to the natural language processing (NLP), which is a computational approach that allows to analyze text that is written in an unstructured mode, as in the case of medical records (MRs). NLP is usually performed in the clinical setting using techniques based on rules given by an expert or a system, but it has been seen that techniques based on machine learning (ML) to increase the performance [12]. In medicine, the NLP has been used in tasks such as extracting relevant information from gastroenterological reports [13], to determine the eligibility of patients for intravenous thrombolytic therapy [14], to manage patients with heart failure from MRs [15], or to support the diagnosis of respiratory diseases from chest X-rays using radiologists reports [16].

However, these approaches are developed with the use of data, which is a problem if the data is unstructured and unavailable. This paper shows the development of a database with clinical reports of suspected TB patients extracted from their MRs. The extracted texts contain information about the patient's health status at times prior to the diagnosis of tuberculosis, with the aim of providing this information to an NLP system that supports the diagnosis of active TB. In addition, some relevant aspects related to specific medical language are provided for the findings in this study.

Andrés Romero is with the Biomedical Engineering Program, Escuela Colombiana de Ingeniería Julio Garavito – Universidad del Rosario, Bogotá D.C., Colombia (e-mail: andres.romero-go@mail.escuelaing.edu.co).

Alvaro D. Orjuela-Cañón is with the School of Medicine and Health Sciences, Universidad del Rosario, Bogotá D.C., Colombia (e-mail: alvaro.orjuela@uroasrio.edu.co).

Andrés L. Jutinico and Erika Vergara are with the Mechanical, Electronics and Biomedical Faculty, Universidad Antonio Nariño, Bogotá D.C., Colombia (e-mail: ajutinico@uan.edu.co, paoli1982@gmail.com).

Carlos Awad and Angélica Palencia are with the Subred Integrada de Servicios de Salud Centro Oriente, Bogotá D.C., Colombia (email: carlosawad@gmail.com, angelicapalenciab@gmail.com).

II. METHODS

A. Dataset Construction

The data were acquired in the *Hospital Santa Clara* from the *Subred Integrada de Servicios de Salud Centro Oriente* in Bogota D.C., Colombia. The *Subred*'s Ethical Review Board approved the project of this work, according to the memorandum number 316 of May 24th, 2021. After collecting data from the period 2017 to 2019, and eliminating inconsistencies, it was possible to collect 151 MRs of patients with TB suspicious. These patients already have the final diagnosis of TB made by specialist using the three different tests given by the diagnosis protocol of Colombia [17]. The employed tests for diagnosis were smear microscopy, culture, and molecular test through the GenXpert©. These analyses allowed to determine 116 confirmed TB cases and 35 patients without the disease. In addition, date information of the tests and the date of treatment initiation were recorded for each patient.

The MRs were stored in Portable Document Format (PDF), according to the public health institution's system. The length of the MRs varies from a few pages to thousands of pages and contains information that may be related to diseases different from TB. Because of this, it was necessary to identify which parts of the MR contain information relevant to the diagnosis of TB, to build the dataset from these sections. Additionally, these extracted sections cannot contain information about the patient's final TB diagnosis, since the idea is to use NLP in suspected TB patients to aid in their diagnosis.

Based on the opinions of the specialists belonging to the project, the information extracted from the MRs consists of reports that physicians make about the patient's health status on the dates prior to testing and initiation of treatment. Also, to have a similar amount of text between patients and to avoid taking records far from the dates of interest, it was decided to include a maximum of five clinical reports and a maximum of 30 days before the treatment´s beginning date.

As the MRs are in PDF format, it was necessary to identify the structure and the form in which the information is presented within the MR, to make a code that automatically reads the documents and extracts the clinical report. Figure 1 shows an example of the format of the MRs, the segment of interest inside of the report, and the date of interest for making decision process of the relevant text extraction.

Figure 1. Example of the format of the MRs.

In order to store and portability of data, different computational techniques were employed through the development of tools in Python programing language. The reports extracted from each patient were saved in a comma separate variables (CSV) files with the respective label, according to the diagnosis of TB given by the specialist. Next, we join all the reports extracted for a patient, what will be called "the patient document". This document was used as the input of the analysis, and the set of all patients is considered as a dataset of labeled text related to TB diagnosis.

B. Dataset Analysis

Before analyzing the content of the dataset, the patients' documents were preprocessed removing Spanish stopwords, numbers, accents, punctuation marks and units of physiological variables. The removal of these items is intended to reduce noise and make the analysis more focused on words with concepts that can provide relevant information regarding TB. Since this is an initial approach to explore the dataset, lemmatization and stemming processes were not considered [14].

Two different strategies were used to visualize the content of the data. The first one is the word cloud, which is visual representation of the words that make up a text, where the size is larger for words that appear more frequently [18]. The other strategy is more complicated and is based on the use of the Term frequency - Inverse document frequency (TF-IDF) measure. TF-IDF take the words in a document and look at their frequency of occurrence (TF), then multiply each one by the inverse of the number of documents in which it appears (IDF) [19]. In this way, words that appear a lot within a document are highlighted, but penalized if they are in several documents. This works according to the possibility that the term appears in several documents, avoiding irrelevant terms. Thus, with TF-IDF it is possible to order the words and highlight those considered most important.

Using TF-IDF, it was decided to keep the top 1000 words with the highest TF-IDF, with a document frequency (DF) in the range of 40-80%, for all the documents. Then for the two groups of patients the occurrence of 2-grams and 4-grams were computed and normalized with respect to the number of patients in each group, to make a comparison between the sets of words that are more frequent in the two groups.

III. RESULTS AND DISCUSSION

The word cloud for the most recurrent words within the dataset is shown in Figure 2. In the image it is possible to observe terms such as *mucosa, oral, dificultad respiratoria, escleras*, which are related to concepts that are used within the framework of the TB. However, words such as *medicina, principal, egreso, horas* are ambiguous and lack real relevance to the disease since they are words that are widely used within the MR to refer to elements of the hospital or to indicate events not related with TB. This was one of the reasons for using TF-IDF, as it allows further refinement of these words to find terms that can really contribute to the differentiation between TB and non-TB patients.

Figure 1. Dataset Word Cloud.

Figures 3a and 3b exhibit the result when the second strategy which involves the TF-IDF measure was applied. These figures show the top 20 n-grams that obtained the highest frequency value with respect to the set of documents. For each n-gram there were two bars representing its frequency value normalized against the number of documents in each class, TB and non-TB patients. In both figures, terms related to expressions with relevance within the framework of TB can be observed, but with better representation that the word cloud. In this case, the n-grams were groups of words that were used in a specific order. For example, *anictéricas conjuntivas normocrómicas* and *mucosa oral húmeda* refer to symptoms of the patients reported by the physicians in the MR. This alternative for text representation allows words to be understood inside of a context, positioning the term into a concept, and not in an isolated mode. Due to this, the n-grams strategy has been used with measures such as TF-IDF in the generation of models that classify documents.

This preliminary results can be employed to physicians of health professionals with low experience in TB, allowing to understand differences in aspects between TB suspicious patients from text reported in medical and clinical records. In this way, the analysis of huge documents can be developed with less time and similar findings. Finally, the present results provide a first approach in the TB diagnosis support problems, where different strategies have been proposed, and where it was possible to find previous effects given by data and the how the information was stored. This allows to determine that the NLP process is a complicated application, especially, when data is unstructured. However, the specialists in pneumology assistance in the development of these systems is still an essential step.

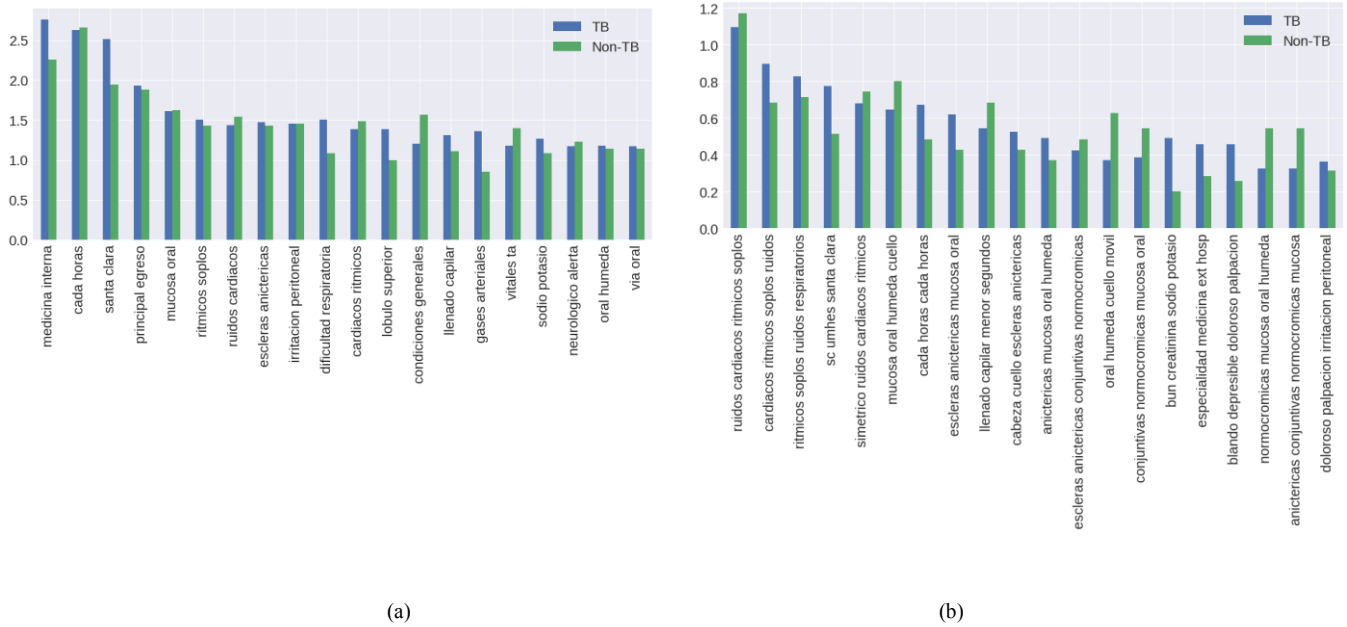
IV. Conclusion

This paper shows the development of a dataset based on clinical reports of suspected TB patients with the aim of facilitating an NLP process. The dataset contains segments of the clinical notes of the physicians in a 30-day period prior to the initiation of the treatment or the first test of TB. To leave the diagnostic task to the NLP models, aspects related to specific dataset were described, and a future work, ML models will be employed to do a TB detection.

The creation of the database was done as part of a larger project that involves the use of other sources of information to support TB diagnosis. It is also hoped that these results can be used in combination with other tools generated within the project to better assist health professionals.

Acknowledgment

Authors acknowledge the support of the Ministerio de Ciencia y Tecnología – Minciencias from Colombia, through founded project 123380762899. In addition, institutions as Universidad Antonio Nariño, the Subred Integrada de Servicios de Salud Centro Oriente and Universidad del Rosario were relevant for the development of this work.



(a) (b)

Figure 1. (a) Most common 2-grams for TB and Non-TB patients. (b) Most common 4-grams for TB and Non-TB patients

REFERENCES

[1] M. A. M. Arcoverde, T. Z. Berra, L. S. Alves, D. T. Dos Santos, A. de Sousa Belchior, A. C. V. Ramos, L. H. Arroyo, I. S. de Assis, J. D. Alves, A. A. R. de Queiroz y others, «How do social-economic differences in urban areas affect tuberculosis mortality in a city in the tri-border region of Brazil, Paraguay and Argentina,» *BMC Public Health,* vol. 18, p. 1–14, 2018.

[2] World Health Organization, «Global tuberculosis report 2020: executive summary,» 2020.

[3] R. Bayer y K. G. Castro, «Tuberculosis elimination in the United States—the need for renewed action,» *New England journal of medicine,* vol. 377, p. 1109–1111, 2017.

[4] Instituto Nacional de Salud, «Informe del Evento Tuberculosis hasta el Periodo Epeidemilógico XI Colombia, 2019,» 2019.

[5] E. Kilsdonk, L. W. Peute y M. W. M. Jaspers, «Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis,» *International journal of medical informatics,* vol. 98, p. 56–64, 2017.

[6] M. Chen y M. Decary, «Artificial intelligence in healthcare: An essential guide for health leaders,» de *Healthcare management forum*, 2020.

[7] S. Houlton, «How artificial intelligence is transforming healthcare,» *Prescriber,* vol. 29, p. 13–17, 2018.

[8] A. D. Orjuela-Cañón, J. E. C. Mendoza, C. E. A. García y E. P. V. Vela, «Tuberculosis diagnosis support analysis for precarious health information systems,» *Computer methods and programs in biomedicine,* vol. 157, p. 11–17, 2018.

[9] Z. Z. Qin, M. S. Sander, B. Rai, C. N. Titahong, S. Sudrungrot, S. N. Laah, L. M. Adhikari, E. J. Carter, L. Puri, A. J. Codlin y others, «Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems,» *Scientific reports,* vol. 9, p. 1–10, 2019.

[10] P. Dande y P. Samant, «Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review,» *Tuberculosis,* vol. 108, p. 1–9, 2018.

[11] S. S. Meraj, R. Yaakob, A. Azman, S. N. M. Rum y A. S. A. Nazri, «Artificial Intelligence in Diagnosing Tuberculosis: A Review,» *International Journal on Advanced Science, Engineering and Information Technology,* vol. 9, p. 81–91, 2019.

[12] S. Doan, M. Conway, T. M. Phuong y L. Ohno-Machado, «Natural language processing in biomedicine: a unified system architecture overview,» *Clinical Bioinformatics,* p. 275–294, 2014.

[13] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug y T. Botsis, «Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review,» *Journal of biomedical informatics,* vol. 73, p. 14–29, 2017.

[14] Y. Wang, J. Luo, S. Hao, H. Xu, A. Y. Shin, B. Jin, R. Liu, X. Deng, L. Wang, L. Zheng y others, «NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records,» *International journal of medical informatics,* vol. 84, p. 1039–1047, 2015.

[15] S.-F. Sung, K. Chen, D. P. Wu, L.-C. Hung, Y.-H. Su y Y.-H. Hu, «Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study,» *International journal of medical informatics,* vol. 112, p. 149–157, 2018.

[16] G. S. Raju, P. J. Lum, R. S. Slack, S. Thirumurthi, P. M. Lynch, E. Miller, B. R. Weston, M. L. Davila, M. S. Bhutani, M. A. Shafi y others, «Natural language processing as an alternative to manual reporting of colonoscopy quality metrics,» *Gastrointestinal endoscopy,* vol. 82, p. 512–519, 2015.

[17] Instituto Nacional de Salud, Tuberculosis: Protocolo de Vigilancia en Salud Pública, Colombia: Instituto Nacional de Salud, 2020.

[18] A. Mueller, «wordcloud: A little word cloud generator,» PyPI, [En línea]. Available: https://github.com/amueller/word_cloud.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot y E. Duchesnay, «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research,* vol. 12, p. 2825–2830, 2011.