Word relevance weighting for the evaluation of medical diagnoses from natural language processing in pulmonary infectious diseases.

Pablo A. Osorio Marulanda[1]

Advanced Multivariate Statistics
Mathematical Engineering
Department of Mathematical Sciences
School of Sciences
Universidad EAFIT

September 2022

[1]paosoriom@eafit.edu.co

1 Problem Statement

Within the dynamics of extraction of information from text that has been explored from the perspective of natural language processing, the application of techniques of this type in the medical field arises, in which the aim is to extract the greatest amount of information from the medical diagnoses that patients receive at the time of being attended, as well as the medical history, both formed from the textual description made by the doctor or expert who attends the patient.

Within the framework of the research carried out by the HUMATH CURIE research group: DECISIONES MEDICAS CONFIABLES EN UNIDADES DE CUIDADO RESPIRATORIO A PARTIR DE INTELIGENCIA ARTIFICIAL EN ENFERMEDADES INFECCIOSAS PULMONARES TIPO NEUMONÍA Y COVID 19, the need arises, based on a set of medical diagnoses describing the medical assessment for patients with pulmonary infectious diseases, to carry out a weighting by words that allow the creation of a score comparable to that already established from the medical perspective for the correct segmentation of patients in terms of disease and severity of it.

The relevance of these indexes in the field of the implementation of Natural Language Processing algorithms is related to the fact that for their establishment sets of notions and terms are used that belong to the development of the description of the clinical case accompanied by the determination of values of laboratory tests and monitoring. In this sense, the establishment of an index, in which the occurrence of these factors determines a weight to establish the patient's mortality risk factor, allows the creation of a search algorithm of terms and values to automatically read the medical records that are in digital repositories and establish the patient's score.

2 Objectives

2.1 General Objective

To create a weighting relevance index that considers the most information from a set of medical diagnoses for the creation of an objective criterion for a patient in the evaluation of pulmonary infectious diseases.

2.2 Specific Objectives

1. To process the data of the medical diagnoses using different NLP techniques.
2. To identify different techniques which have been used for a similar purpose through a literature review of the main topic.
3. To create a machine learning model to index the words according to their importance inside the text and beside the label of the diagnostic for the pulmonary infection disease.
4. To compare results with the scores generated from a medical perspective.

3 Proposed Methodology

Cross-Industry Standard Process for Data Mining (CRISP-DM) .[2], CRISP-DM is a widely-used analytical model that aids in the planning, organization, and implementation of data projects. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Implementation are the six processes that make up the CRISP-DM approach. The following is a description of these phases:
Business understanding: In this phase, the project's objectives and needs are understood, the general and particular goals are established, the situation is evaluated, and the project plan is created.
Data understanding: Drives the emphasis to describe, explore, and confirm data quality in order to identify, gather, and analyze the data set that will be used.
Data preparation: The sample will be chosen, the dataset will be cleaned, the dataset will be constructed, and if required, the data will be reformatted to create the final data set.
Modeling: The model or models that will be employed to address the project's question must now be chosen. This section requires you to decide which models to test, develop a set of tests for the models, build the models, and then analyze each model's output.
Evaluation: The findings from each model should now be compared, evaluated, and checked to see whether they all met your requirements. Once you've selected the model or models to use, the following step should be decided.
Implementation: The model's usefulness will be completed once it is deployed and implemented, for that testing is of utmost importance so issues during the operational phase can be avoided. The final report will be produced in this period.

For this project, we will consider some of the techniques discussed in Rathi & Mustafi (2022), which propose the review of a set of techniques such as Binary weighting, Term frequency, Term frequency-inverse document frequency, Supervised term weighting, Confidence weight, Unsupervised term weighting, Feature-based term weighting, among others.
Other explorations with a high value in our project objective can be to consider the Skip-gram Model as is exposed in Mikolov *et al.* (2013), as well as to explore the possibilities with a word2vec model.

4 Data Description

The data that will be used for this model is a collection of medical diagnoses collected from descriptions of patients with respiratory diseases, which have 3 labels: covid, normal, and other. The balance of the classes can be visualized in Figure (1). The data are already preprocessed, while the diagnoses are now only represented with stemming and stopwords. There are 844 registers of the data. Considering medical files, it is possible to generate a vector space model, or term vector model, which is an algebraic model for representing text documents as vectors, where each dimension represents the frequency of either stemming

[2]IBM. CRISP-DM Help Overview. `https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview`

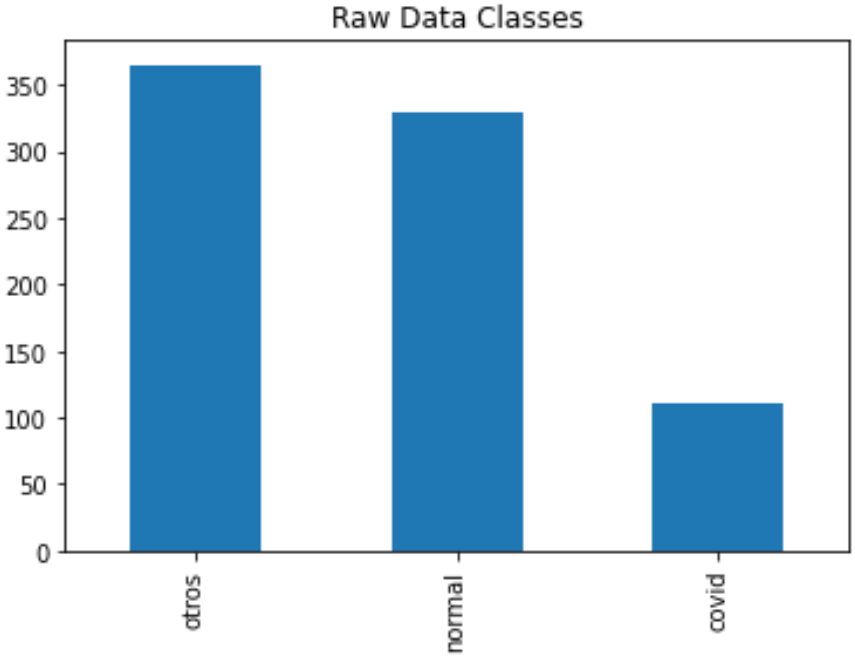

Figure 1: Bar plot of the label distribution for the raw data

words or stopwords resulting from tokenization by words and grammatical roots. Using Principal Component Analysis, it is possible to observe the representation of this space, which is observable in the figures (2, 3)

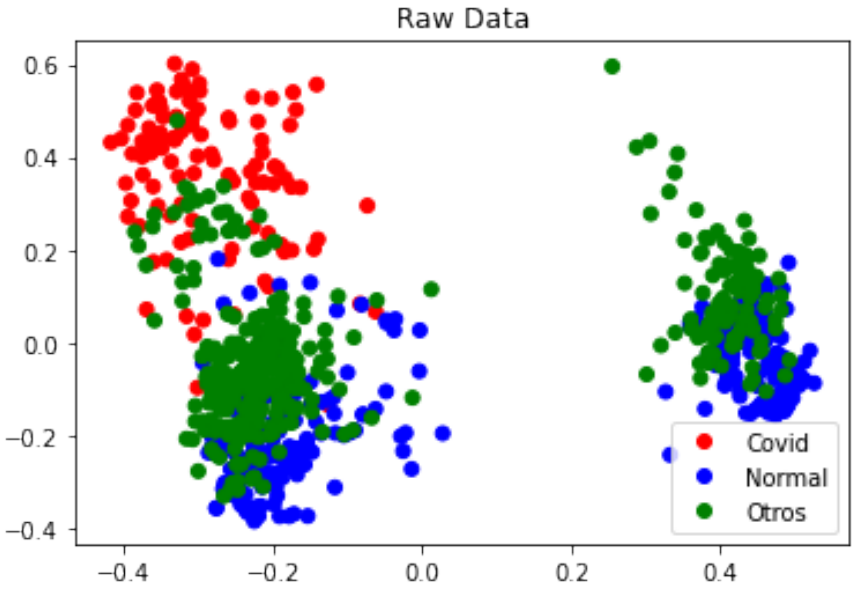

Figure 2: Stopwords frequency Vector Space Model using PCA

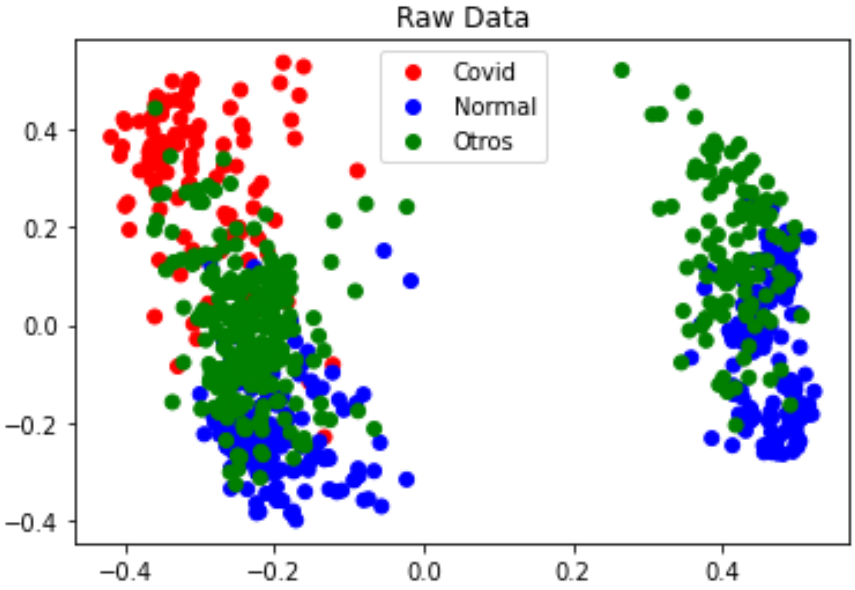

Figure 3: Stemming frequency Vector Space Model using PCA

The dimension of the Stopwords Frequency Matrix is 806 rows × 125 columns and the Stemming Words Frequency Matrix is 806 rows × 131 columns

5 Schedule

Activity	Weeks												
	6	7	8	9	10	11	12	13	14	15	16	17	18
Bussiness understanding	■	■	■	■									
Data understanding			■	■	■								
Data preparation				■	■	■							
Modeling						■	■	■	■				
Evaluation							■	■	■	■			
Implementation						■	■	■	■	■	■	■	
Final document Writting											■	■	■

Figure 4: Schedule

6 Ethical Implications

Considering that we are working with medical records, there are several ethical considerations to take into account. The first of these falls on the action of keeping the data secure, in that there must be direct actions for the security and confidentiality of patient data. The second has to do with the ability to keep the identity of patients confidential, and for some cases completely anonymous, as a certain amount of data associated with an individual creates

additional opportunities to identify them - permitting deductive identification, and this may lead to particular actions being taken against the individual, which, for some cases may be beneficial in terms of public health, but in others, may bring disadvantages for the individual in a marginal way, which brings us to the third point: Basing action on data and evidence. Data considerations such as data quality, under-representation of data on complex issues, conflicting evidence, and lack of data as an excuse to postpone action must be taken into account when making real decisions, as these conflicts can lead to an inappropriate analysis of health and service patterns and trends, and misappropriation of resources.

7 Commercial and Legal Aspects

From the legal perspective, there are considerations regarding the use of the data, since they must be used confidentially, treating patients globally, as a statistic and not in a particular way. On the other hand, the commercial aspect that, even remote for the considerations of a project, should consider this type of work can be subsequently employed by health institutions as a tool of augmented intelligence for the medical analysis of diagnoses, for support before the treatments to be employed.

References

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. 2013. *Distributed Representations of Words and Phrases and their Compositionality.*

Rathi, R. N., & Mustafi, A. 2022. The importance of Term Weighting in semantic understanding of text: A review of techniques. *Multimedia Tools and Applications*, Apr.