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: Reliable medical decisions in respiratory care units based on artificial intelligence in pulmonary infectious diseases such as pneumonia and 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 Literature review

Since according to Salazar Martínez *et al.* ((s.f)) NLP has been utilized in situations where access to a lot of written texts makes it possible to find patterns relating to how semantic fields are constructed, it is relatively new but widely used to identify and classify imaging results, prioritize patients, create imaging procedures, and extract data of research relevance.

One of the NLP main research areas is text mining. The text mining process involves eliminating irrelevant information from unstructured data, extracting it as structured data, and then using data mining algorithms to find the important information. However, it is crucial to apply filtering procedures to lower the space dimensionality because it is typically a major issue in statistical text clustering. One of the main methods for this is the stemming algorithms, which relate morphologically similar indexed and search terms, it is used to improve the retrieval effectiveness, add some robustness to the statistical frequency representation of the documents and reduce the dimensionality of the problem Montoya *et al.* (2015).

Works such as Salazar Martínez *et al.* ((s.f)), Vidal-Correa *et al.* (2022), Spyns (1996a) explore through NLP the incidence, influence and contribution from a medical evaluation perspective. Salazar Martínez *et al.* ((s.f)) proposes the use of a strategy to establish the type of statements and notions that are common to lung disease diagnoses, while Vidal-Correa *et al.* (2022) uncovers the internal structure of the data to identify the most informative data in a set of papers using 44K+ papers with medical findings for COVID-19. Spyns (1996a) performs a state-of-the-art review of the NLP domain in the medical field, which concludes the broad contribution of NLP to the medical field considering Especially in this field, where many concepts are language-independent, very good results are obtained by domain-dependent language analyzers - especially concerning grammatical input.

4 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.

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

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.

5 Methods

In this section, some of the used methods are going to be described.

5.1 Visualizations

5.1.1 Textnets

The Textnet technique captures both internal hierarchical structures and graphlike text pools using a single uniform data structure. The associations between nearby bits of text are made clear by employing a semantic network formalism of nodes connected by typed links. A semantic net formalism of labeled connections and nodes is used to organize the Textnet network. Chunks and Tots, which represent the textual and hierarchical components, respectively, are the two different sorts of nodes. As a result, a chunk node represents a basic paragraph of text, but a tot node roughly equates to a table of contents item. T,yped (labeled) linkages connect the different nodes. The type of the link is intended to describe how two nodes are related to one another (Trigg & Weiser, 1986).

5.1.2 WordClouds

Word clouds created for a body of text can be used as a jumping off point for more in-depth investigation. For instance, they assist in determining if a text is pertinent to a certain information demand. One of their shortcomings is that they don't take linguistic understanding of the words and their relationships into account, providing only a statistical summary of isolated words. As a result, most systems employ word clouds fairly statically to summarize text, and they often provide no or very limited interactivity options (Heimerl *et al.*, 2014).

5.1.3 TSNE

T-SNE is a technique for high-dimensional data visualization by giving each data point a location in a two or three-dimensional map that can use random walks on neighborhood graphs to allow the implicit structure of all of the data to influence the way in which a subset of the data is displayed, using a Stochastic Neighbor Embedding (Van der Maaten & Hinton, 2008).

5.2 Vectorizers

5.2.1 TF-IDF

The TfidfVectorizer calculates the appropriate term frequency for each word using the inverse domain frequency (idf) and term frequency (tf) of the words. frequency in the inverse domain values for (Tf-idf). When completing the sentiment analysis, the (Tf-idf) value is then utilized to provide the word weight or significance. Using the equations shown below, the idf of the word "w" in the text corpus and the Tf-idf of the word "w" in a specific document d are calculated, where, $tf(w, d)$ is the term frequency of word 'w' in document d, which denotes the number of times 'w' appeared in document d divided by total number of words in the document (Subba & Gupta, 2021).

$$idf(w) = log\left(\frac{\text{total number of docs}}{N.ofdocswithwordw}\right) \quad TF - df(w, d) = tf(w, d) * idf(w) \tag{1}$$

5.3 Feature selection

5.3.1 Chi-Square and Mutual Information Classification

Among the feature selection techniques, there are the ones called "filter feature selection methods", which calculate some stats from the data and give every feature a score to select the best k (Chandrashekar & Sahin, 2014). For this work, we choose two of them:

- **Chi-square**
  A chi-square stat measures how expected count E (expected value) and observed count O (observed value) deviates each other with c degrees of freedom.

  $$\chi_c^2 = \sum \frac{(O_i - E_i^2)}{E_i} \tag{2}$$

  Consider a situation where we need to figure out how the independent category feature (predictor) and dependent category feature relate to one another (response). When choosing features, our goal is to choose those that depend heavily on the outcome.

  The observed count is close to the anticipated count when two characteristics are independent, hence the Chi-Square value will be lower. Therefore, a high Chi-Square score suggests that the independence hypothesis is false. Simply said, features that are more reliant on the response and have larger Chi-Square values might be chosen for model training.

- **Mutual information**
  Mutual information ranking criteria use the measure of dependency between two variables. To describe it is necessary to calculate Shanon's entropy

$$H(Y) = -\sum_{y} p(y)log(p(y))$$

  where H represents the uncertainty in output Y. The conditional entropy is described by

$$H(Y|X) = -\sum_{x}\sum_{y} p(x,y)log(p(y|x)) \tag{3}$$

  The conditional entropy implies that by observing a variable X, the uncertainty in the output Y is reduced. The decrease in the uncertainty is given as

$$I(Y,X) = H(Y) - H(Y|X) \tag{4}$$

  I gives the Mutual Information between Y and X, meaning that if X and Y are independent, then MI will be zero and greater than zero if they are dependent.

5.3.2 Random forest and Logistic Regression

For implementing RF and LR as regressors to solve the problem of selecting the best features we implement the wrapper methods algorithm. Wrapper methods evaluate the variable subset using the predictor performance as the objective function and the predictor as a black box. Due to the difficulty of evaluating 2^N subsets, suboptimal subsets are discovered by using search algorithms that heuristically find a subset.

6 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 words or stopwords resulting from tokenization by words and grammatical roots. An example of a typical register is shown in figure 2

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

7 Results

According to the methods discussed in the section of methods, here are some of the founded results

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

text:{ "hospital universitario san vicente fundacion paciente gonzalez martinez fany carmen servicio u.e uce infa santiago corazon solicitud estudio rayos x torax portatil id cc informado -may- : pm realizado -may- : pm indicaciones diarrea hallazgos traquea posicion anatomica aumento indice cardiotoracico observa ensanchamiento mediastino opacidades alveolares bilaterales sugiere neumonia multi lobar signos derrame pleural neumotorax electrodos externos monitoreo osteopenia difusa attentamente informado ruiz zabaleta tania isabel md radiologo rm:- hospital universitario san vicente fundacion medellin calle d- medellin telefonos . extensiones:,, correo electronico imaginologia sanvicentefundacion.com "}
type: {covid}

Figure 2: A register of the data

7.1 Visualizations

Using a textnet it is possible to visualize not only a clear cluster structure between documents, but also a clear relationship of concepts between them. The figure (3) shows it properly

Additionally, considering that what we are looking for are the relevant words per class, it is appropriate to make a visualization in wordcloud of the most frequent words per class. This is possible to observe in the figure (4). Accordingly, no unknown linguistic structure in terms of description of the described lung disease is shown, as it is only possible to see typical words in this type of diagnosis.

7.2 Proyection of VSM

Now considering a counting vectorizer (frequency of words in the text) and a TFIDF, in addition to using the projections of the vector space generated by each of these methods through a TSNE, the detail is as follows:

A clear set of groupings can be seen in the figures (5, 6) with regard to covid-labeled

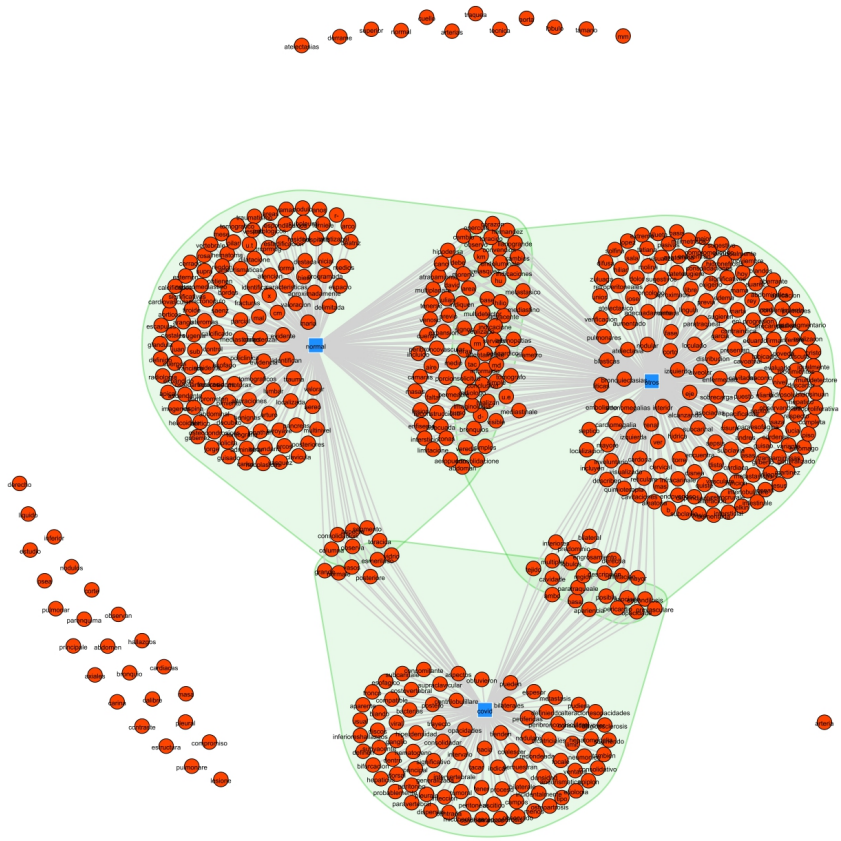

Figure 3: Textnet structure of the data

Figure 4: Wordcloud per class

diagnoses. For the category "other" and "normal", the cluster is not observably separable.

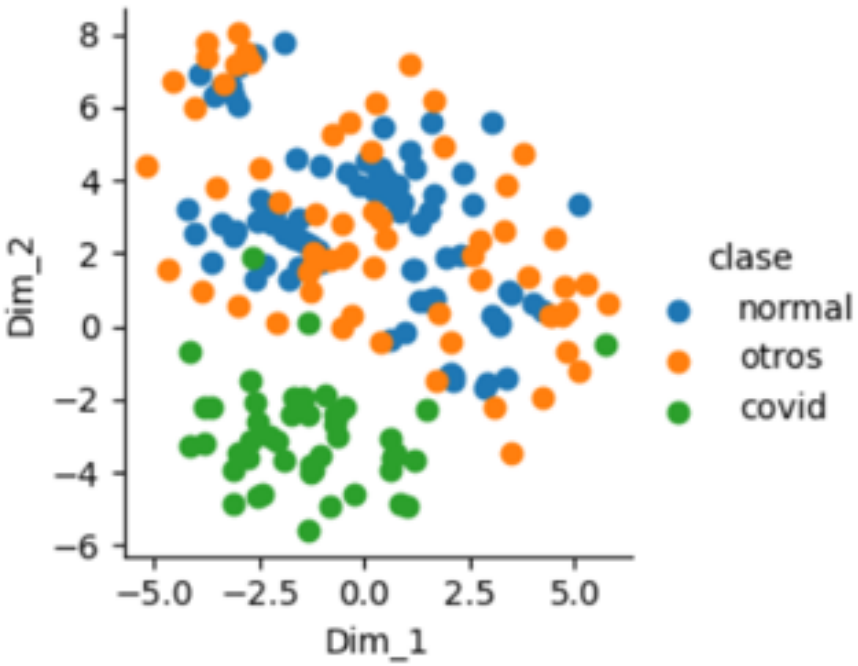

Figure 5: TSNE protection with a count vectorizer

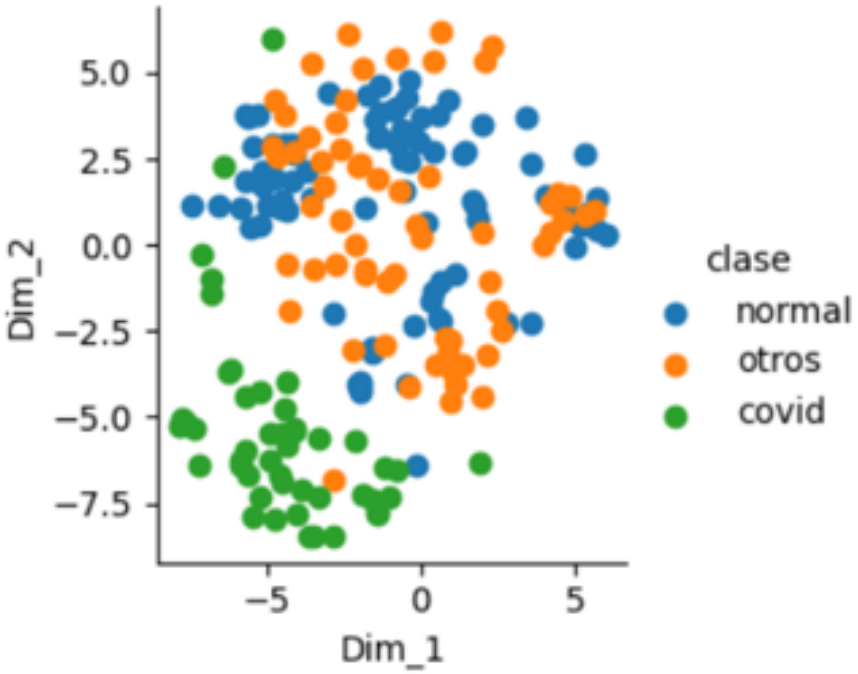

Figure 6: TSNE protection with a TF-IDF vectorizer

	chi_square	mutual_info_classif	Random_Forest	Logistic_Regression
0	observando	servicio	bilateral	anormales
1	viral	informado	enfisema	attentamente
2	informado	radiologo	informado	bilateral
3	multicorte	conclusion	lobulo	informado
4	realizo	solicitud	masas	md
5	cantidad	realizado	pleural	multiples
6	bilateral	attentamente	radiologo	patologicos
7	tronco	indicaciones	realizado	policlinica
8	patron	md	servicio	servicio
9	cica	rm	solicitud	viral

Table 1: Top 10 features for VSM count vectorizer using Feature Selection

	chi_square	mutual_info_classif	Random_Forest	Logistic_Regression
0	observando	servicio	bilateral	bilateral
1	viral	radiologo	estudio	cantidad
2	multicorte	informado	informado	cuello
3	cica	realizado	masas	informado
4	tora	solicitud	pleural	masas
5	realizo	conclusion	radiologo	multiples
6	espesor	md	realizado	normal
7	intervalo	attentamente	rm	observando
8	obtuvieron	indicaciones	servicio	servicio
9	covid	rm	solicitud	viral

Table 2: Top 10 features for VSM using TFIDF vectorizer using Feature Selection

7.3 Feature selection results

Considering that we have two different VSMs, we can evaluate the described methods for feature selection in both, and thus see which word vector is the most representative in terms of the described statistics and predictions with the implemented models. The following table represents the top 10 most important characteristics for the VSM of the count vectorizer and TFIDF.

8 Conclusions and Future Work

While the results of the score achieved through feature selection are useful in that the best words for prediction are known, and it allows us to achieve an importance score for the words, which was one of the initial objectives of the research. However, one of the ultimate goals

of developing these mechanics is to develop methods that allow us to see the location of the lung condition, and with the methods implemented here, it is not possible to observe this. Additionally, one of the limitations of using feature selection algorithms is not knowing to what extent a certain word influences the diagnosis by class. This is why as future work we propose not only the implementation of more optimal feature selection models, but also more robust in terms of information gain, such as the respective implementations in BERTO or other NLP-heavy tools.

9 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 7: Schedule

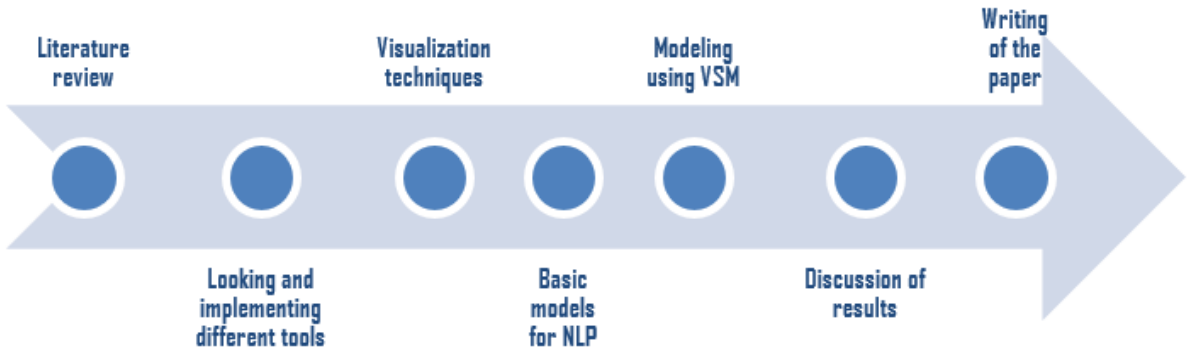

Figure 8: New plan

According to the diagram 8, it is possible to observe that the parallelization of tasks is somewhat complicated in this type of process since you generally wait to finish a task to start the next one. The schedule used was very ambitious in terms of time and objectives, however, the objectives were met to the respective extent.

10 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.

11 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

Chandrashekar, Girish, & Sahin, Ferat. 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, **40**(1), 16–28.

Heimerl, Florian, Lohmann, Steffen, Lange, Simon, & Ertl, Thomas. 2014. Word cloud explorer: Text analytics based on word clouds. *Pages 1833–1842 of: 2014 47th Hawaii international conference on system sciences*. IEEE.

Hernández, Myriam Beatriz, & Gómez, José M. 2013. Aplicaciones de procesamiento de lenguaje natural. *Revista Politécnica*, **32**.

Kraskov, Alexander, Stögbauer, Harald, & Grassberger, Peter. 2011. Erratum: estimating mutual information [Phys. Rev. E 69, 066138 (2004)]. *Physical Review E*, **83**(1), 019903.

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

Montoya, O Lucia Quintero, Villa, Luisa F, Muñoz, Santiago, Arenas, Ana C Ruiz, & Bastidas, Manuela. 2015. Information retrieval on documents methodology based on entropy filtering methodologies. *International Journal of Business Intelligence and Data Mining*, **10**(3), 280–296.

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.

Salazar Martínez, C.A., Francisco Vargas, J., Hernandez Arango, A., Garcés, J.J., Ramírez Cadavid, D. A., & Quintero, O. L. (s.f). Is Natural Language Processing a suitable tool for intelligent systems in radiology? a Spanish Corpus of Descriptions of Radiological Findings in the Framework of the SARS-CoV-2 Pandemic. *Revista Politécnica*, **32**.

Spyns, Peter. 1996a. Natural language processing in medicine: an overview. *Methods of information in medicine*, **35**(04/05), 285–301.

Spyns, Peter. 1996b. Natural language processing in medicine: an overview. *Methods of information in medicine*, **35**(04/05), 285–301.

Subba, Basant, & Gupta, Prakriti. 2021. A tfidfvectorizer and singular value decomposition based host intrusion detection system framework for detecting anomalous system processes. *Computers & Security*, **100**, 102084.

Trigg, Randall H, & Weiser, Mark. 1986. TEXTNET: A network-based approach to text handling. *ACM Transactions on Information Systems (TOIS)*, **4**(1), 1–23.

Van der Maaten, Laurens, & Hinton, Geoffrey. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, **9**(11).

Vidal-Correa, J.P, Salazar Martínez, C.A, Posada, A.Mejía Camila, Ariza-Jiménez, L., Murillo-González, A., Garcés E., J.J., Quintero, O. L., Guillermo Paniagua, J., Arrazola, W., Ramirez Cadavid, D.C., & Restrepo, T. 2022. Information Retrieval from NLP during COVID-19 for Treatment in Developing Countries. *International Journal of Business Intelligence and Data Mining*.