Using Latent Semantic Indexing as a metric for evaluating research potentialities through Innovation Public Policies

Renan Oliveira Silva[1] and Rafael Lima de Carvalho[1]

[1] *Universidade Federal do Tocantins, Computer Science Department, Tocantins, Brazil*

Reception date of the manuscript: 26/06/2021
Acceptance date of the manuscript: 02/08/2021
Publication date: 10/08/2021

Abstract— Public innovation policies usually define strategies for public research organizations, such as universities, in order to guide the next research projects of such organizations. Sometimes, it is difficult to know the actual state of an organization when a new policy is released by the government. The objective of this paper is to present the application of Latent Semantic Analysis, a technique of information retrieval, in order to create an index and automatically classify research projects, using text fields like title and abstract, to areas and subareas defined by related terms. It is also proposed a case study of about 200 projects from five graduate programs of the *Universidade Federal do Tocantins*. The proposed solution was capable of satisfactorily classify each project to the areas and subareas of a recent policy from the Science, Technology, Innovations, and Communications Ministry. In this way, the university could have some decision-making information, and the results could sustain for which internal policies could be implemented to maximize its actuation faced to the national innovation policy.

Keywords—Latent Semantic Analysis, Science and Technology, Research and Innovation Policies.

I. Introduction

Throughout history, humanity has been continuously evolving as a society. We have been learning over the centuries to form the concept of our "modern society". This concept, when applied acceptably, can give certain rights to the people and should develop laws or policies attempting to provide services to the social well-being. Such policies, both public and private, need to be evaluated because the implementation of these policies relies on investments. In the public context, a policy is drafted aiming to address a solution to a certain public problem. A policy is commonly associated with a program and there are expectations that it will fix a problem, which implies that there is something to be fixed, i.e., a problem [1].

The policies evaluation can occur in different moments within the policy cycle, each one is used for different purposes, according to when it is performed [2]:

- ***ex-ante***: it happens at the beginning of the policy creation process, though, it provides strategical information that determines the continuation of a policy.
- **interim**: it lies sometime between the *ex ante* and the *ex post* evaluations. It can help avoid or correct mistakes within the development process. Moreover, it assesses the results of the implementation phase, providing control information.
- ***ex post***: it is performed at the end of the policy cycle, verifying the policy impacts. Furthermore, it may help in design and decision-making on similar projects in the future.

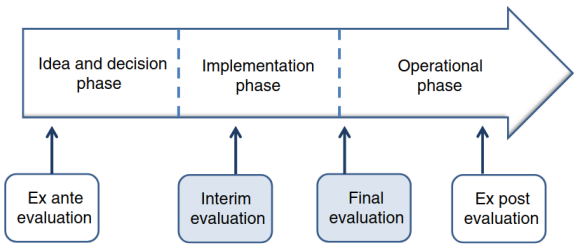

Fig. 1: The policy evaluation in different moments.

The Figure 1 [2] illustrates the policy cycle evaluation timeline, highlighting the policy evaluation in different moments. According to [3], matching the requirements of policymakers with the skills and experience of evaluators can reveal crucial divergences in perspectives. Likewise, such divergences may affect the delivery and implementation of evaluation studies that serve the desired policy purposes. Additionally, the policies arise, interact and have effects in a messy, complex, multi-level, and multi-actor reality [4]. The Figure 2 [3] illustrates the complexity of defining a metric for the evaluation study.

Contact data: Renan Oliveira Silva, renan.oliveira@mail.uft.edu.br

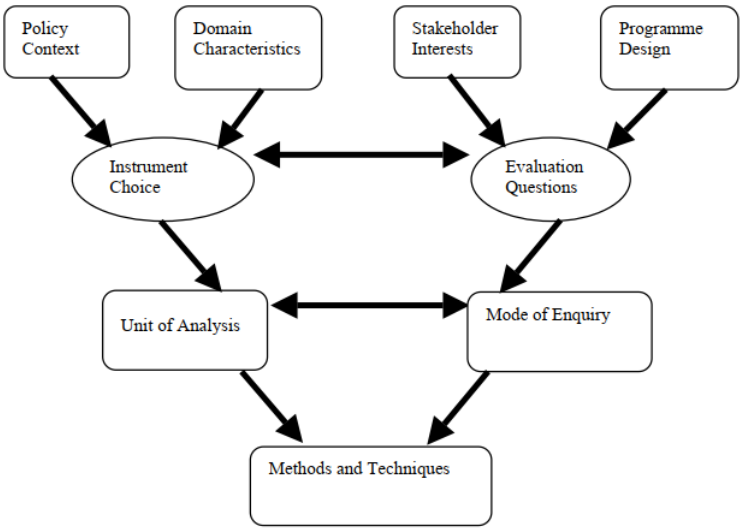

Fig. 2: A framework for formulating an appropriate evaluation methodology.

In Brazil, most of the public policies concerning research and development are provided by the MCTIC (*Ministério da Ciência, Tecnologia, Inovações e Comunicações*[1]). In 2020, MCTIC released an ordinance defining priority areas with regard to research projects in national scope, for the development of technologies and innovations for the period from 2020 to 2023 [5]. In such document, they have pointed five great strategic areas known as: Strategical Technologies, Skilled Technologies, Production Technologies, Technologies for Sustainable Development and Technologies for Quality of Life. Each area is divided into more specific subareas, defining the strategies for the development of further research projects.

Many Brazilian organizations are responsible for the development of research projects. Much of such production comes from public universities [6], which have many professors actuating as researchers. Usually, these researchers register their projects using the Lattes Platform[2] (a national platform used to store curriculum data for researchers). So, to know which projects the researchers are working on in a period, the Lattes data can be used, but textual information should be processed to give a semantic evaluation of such data. Therefore, it could be time-consuming to evaluate the potentialities of such research organizations, when an innovation policy is released.

In this work, we propose the use of Latent Semantic Indexing to automatically associate projects of a research organization into an innovation policy formed by areas and subareas, with free terms classifying each subarea. The research question is: given the representation of research, such as documents consisting of title and abstract, and an innovation policy, which describes terms of areas and subareas, is it possible to use LSI to automatically classify each project into the policy to evaluate the current potential of innovation of such organization?

This document is organized as follows: Section II describes the necessary background covering the main technical subjects such as LSI and correlated textual tools. In Section III we discuss, through some examples, the LSI effectiveness in different applications. In addition, in Section IV we present the adopted methodology, which defines the sample database, scope and delimitation, as well as the main pipeline connecting the textual tools. Further, Section V presents the analysis of the results when applying the methodology to a Case Study of a University and the Brazilian Science and Technology ministry's innovation policy for the country. Finally, Section VI shows some final remarks of this research as well as some future insights.

[1]Science, Technology, Innovations and Communications Ministry
[2]http://lattes.cnpq.br/

II. Background

In this paper, we use Latent Semantic Indexing (LSI) as a policy evaluation metric. The LSI, also referred to as Latent Semantic Analysis (LSA), is a technique for information retrieval that takes advantage of implicit structures and recognizes semantic relations between terms and documents [7]. This approach also produces more concise results in comparison with matching term approaches.

a. Singular value decomposition

The core of the LSI is the Singular Value Decomposition (SVD). As the name implies, the SVD decomposes the term-document matrix, which will be discussed later, and transforms it into a lower-dimensional matrix. At this point, the purpose is to deal with non-square and probably non-symmetric matrices [8]. Applying the SVD, the term-document matrix A_{txd} will be decomposed into the product of T_{txn}, S_{nxn}, and D_{dxn}:

$$A_{txd} = T_{txn} S_{nxn} (D_{dxn})^T$$

where t is the number of terms, d the number of documents and $n = \min(t, d)$ [9]. The columns of T and D are orthonormal, it means that $TT^T = D^T D = I$. Furthermore, the S is a diagonal matrix in the descending order containing the eigenvalues of $A_{txd}(A_{txd})^T$ that are the same for $(A_{txd})^T A_{txd}$. In the SVD, the eigenvalues are known as *single values*. The restriction of T, S and D to their first $k < n$ rows, results in

the T_{txk}, S_{kxk} and D_{dxk} matrices [9]. The product of these matrices is $\hat{A}$, which is:

$$\hat{A}_{txk} = T_{txk} S_{kxk} (D_{dxk})^T$$

and it is the best square approximation of A by a matrix of rank k as defined in the equation $\Delta = ||A - \hat{A}||$ [9]. We are not going to discuss SVD deeply in this paper, but a brief description is that the decomposition allows us to represent a huge matrix (term-document matrix) in a lower dimension. That is why $k < n$ and the $\hat{A}$ matrix has a lower rank than A. Additionally, the $\hat{A}$ matrix represents the synonyms more precisely than A [9].

b. Preprocessing

Before applying the LSI in the input data, it is necessary to take out unwanted parts of the dataset in a process known as preprocessing. At the beginning of this process, the stopwords are removed. The stopwords show up frequently in natural language documents, but it provides minimal contextual information [10]. Example stopwords include each, about, such, and the. The next step is to reduce words with different grammatical inflections. In such a process, there are two main approaches: stemming and lemmatization. Both techniques reduce words into a base form, but they work in a slightly different manner.

The stemming consists of converting morphological forms of words to their stem, a stem does not need to be an existing word in the dictionary, but all its variants should map to this form after the stemming process [11]. On the other hand, lemmatization transforms a word into a lemma: a canonical form of a lexeme. The lexeme is a set of all word inflections that have the same meaning. The stem is formed by plenty of rules used in the stemmer algorithm, while the lemma is chosen by convention to represent a certain lexeme.

c. Dictionary and bag of words

After the preprocessing, all the documents within the dataset are analyzed and the documents' words are stored in a dictionary. That dictionary is a lookup table containing the document frequency of a term as well as where in the postings file the per-document counts can be accessed [12]. The postings file is an inverted file with the per-document counts and the IDs of the documents, as well as the position of each term occurrence.

Thus, the bag of words (BoW) is created and, according to [13], it is a simple representation of text that is used in retrieval and classification models. In this representation, a document is considered to be an unordered collection of words with neither syntactic nor statistical relationships between them. Then, all documents are "reshaped" according to the BoW, i.e., the documents are represented by the term occurrences. At this point, the purpose is to arrange the documents as vectors in the vector space model.

d. Document-term matrix and vector space model

According to [13], in the vector space model, documents and queries are assumed to be part of a t-dimensional vector space, where t is the number of index terms and a document is represented by a vector of index terms as follows:

$$D_i = (d_{i1}, d_{i2}, ..., d_{it}),$$

A document collection containing n documents (document-term matrix) of term weights, each row representing a document and each column describing the weights assigned to a term for a certain document, can be represented like this:

	$Term_1$	$Term_2$	...	$Term_t$
$Document_1$	d_{11}	d_{12}	...	d_{1t}
$Document_2$	d_{21}	d_{22}	...	d_{2t}
⋮	⋮	⋮		⋮
$Document_n$	d_{n1}	d_{n2}	...	d_{nt}

Because the documents are vectors to be arranged in a vector space [12], the Figure 3 illustrates a vector space with two coordinates (terms).

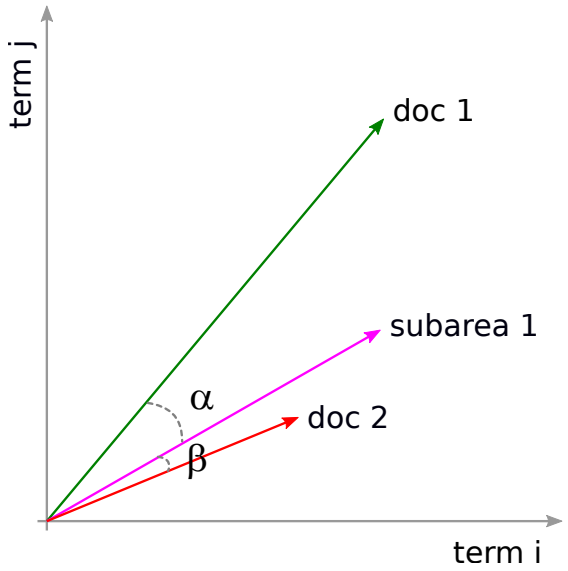

Fig. 3: Documents plotted in a vector space with two coordinates.

As vectors plotted in a vector space, the documents can be compared to determine the similarities between them. Both query and documents are vectors plotted in the vector space, consequently, they are pointed and there are angles among them [12]. One approach broadly used to compute such angles is the cosine similarity. The documents are ranked by computing the similarity between the query and each document in the document collection.

When plotting the documents in the vector space, the model ranks the documents according to a number of topics, which varies depending on the dataset. A common practice is to compute the ideal number of topics for each dataset, given an interval. For example, if the interval is $[1, 100]$, an algorithm will calculate the topic coherence for each number within this interval. The number of topics with the greatest coherence score will be used in the model to rank the dataset. After the ranking, the model utilizes a threshold given by the user, if a document score is less than the threshold, it will not be shown in the ranking results.

e. TF-IDF

As previously mentioned, the vectors are pointed and because their coordinates are terms, each coordinate in a vector is a weight that points the vector. At first, a weight represents the presence of terms in a document, because of that, the vectors assume two values: zero or one. When a term is present in a document, it is assigned one to the term coordinate, otherwise, it is assigned zero. But considering just the presence or absence of a word is a naive approach, that becomes clear when realizing that the presence of a term in a document does not tell us anything about the documents context [14].

A more precise approach is assigning weights as term per-document counts, which gives us more refined results. But to improve the ranking of the documents even more, we need to consider some constraints. Using the term per-document count directly as the vector weight may be accurate enough, mainly if we consider very frequent words in a document. That is why, documents with high term frequencies can be ranked over documents with a more interesting context, for example. Another point is that very frequent words throughout the collection do not have the same relevance as rare words when distinguishing a document context.

Aiming to address a solution for the problems described above, the TF-IDF is a commonly used function for assigning the vector weights. The TF-IDF consists of both term frequency (TF) and inverse document frequency (IDF), while the TF value considers directly the term occurrence, the IDF penalizes very common terms throughout the document collection and rewards rare words [12].

III. Related Works

Since its formulation in the 1980s [15], several researches have applied LSI. A well-known application is in search engines, which uses the LSI power to classify a massive amount of stored documents, which became globally used after the internet break out in the 1990s. The search engines utilize information retrieval techniques, such as LSI, to rank millions of documents according to a user's query.

Besides the query-search usage, the LSI can be applied for different purposes. Some researches like [16], [17], and [18] utilize it to approach different problems. Even if [19] research aims at a different objective in comparison to ours, both apply the LSI in a quite similar manner, classifying papers according to title and abstract. Moreover, some works combined different techniques to enhance the LSI power, as seen in [20] and [21].

IV. Methodology

a. Summary

Primarily, it was necessary to create a dataset to be used when applying the LSI. A research was made by [22] and the data refers to the works of the UFT researchers and the data sources was both MCTIC (*Ministério da Ciência, Tecnologia, Inovações e Comunicações*) and the *Currículo Lattes*. Then, the research results were used to mount a dataset which is used as input by this work.

At this point, the given input is refined in the preprocessing and, lately, plotted in the vector space model. As discussed previously, the dataset documents (papers) and the queries (pseudo-documents) are vectors in the vector space. Now, the model computes the similarity between each query and the documents according to a threshold and a number of topics. Lastly, the ranking results are plotted in a graph.

b. Scope and delimitation

The first data source is composed of two main documents: a) "ENCTI 2016-2020", which is a document with the Brazilian's national strategy for Science, Technology and Innovation for the time interval between 2016 and 2020; and b) MCTIC Ordinance No. 1,122/2020 (later partially amended by the MCTIC Ordinance No. 1,329/2020), which defines the priority research areas of the national strategy, covering the period from 2020 to 2023. Each area is separated into sub-areas or sectors, namely:

- **Strategical Technologies**: Spatial; Nuclear; Cybernetics; and Border and Public Safety;
- **Skilled Technologies**: Artificial Intelligence; Internet of Things; Advanced Materials; Biotechnology; e Nanotechnology;
- **Production Technologies**: Industry; Agribusiness; Communications; Infrastructure; and Services;
- **Technologies for Sustainable Development**: Smart Cities; Renewable Energies; Bioeconomy; Solid Waste Treatment and Recycling; Pollution Treatment; Monitoring, prevention and recovery from natural disasters and environmental; and Environmental Preservation; and
- **Technologies for Quality of Life**: Health; Sanitation; Water Security; and Assistive Technologies.

This research is applied to a Case Study of the Federal University of Tocantins (UFT). Currently, the UFT has 33 graduate programs. Based on the aforementioned areas, this research considered a sample composed of the following graduate programs: *Biotecnologia*, *Agroenergia*, BEC, CTA, MCS and Profnit.

Such a sample includes the name of the researchers and their projects within the graduate programs. All the information about the projects was collected and there are 95 researchers in this sample. To take the most recent data, it was considered only active projects from 2018 to 2020. The projects' data source was the *Currículo Lattes*, which is a governmental platform used to update researchers' curricula. From that point of view, it has been collected an amount of 211 projects, which contains title and abstract.

The pseudo-documents used are collections of terms representing each subarea of the MCTIC. The terms composing the collections are common terms shared by projects located in the same subarea. The pseudo-documents were composed by [22] and it was not considered any constraints when eliciting such terms, the only concern was selecting terms that were contextually important to the subarea and could, as better as possible, describe it with the source of [23].

c. Flowchart

The flowchart in Figure 4 demonstrates the whole evaluation process and, as already stated, a dataset is a collection of pseudo-documents characterizing the MCTIC's subareas. Therefore, the input is prepared to be analyzed and ranked by the LSI, which computes the similarity between queries and documents. After the entire process, the results are shown in a graph.

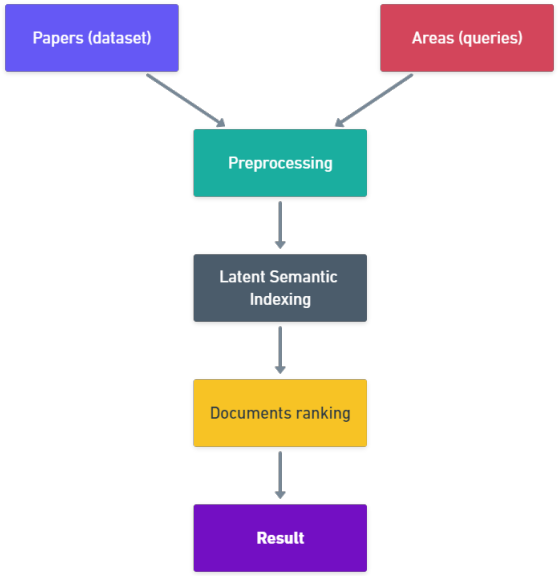

Fig. 4: The evaluation process flowchart.

V. Results

The evaluation process is based on analyzing the similarity of each project in the dataset with each given subarea. Each subarea of the MCTIC belongs to a great area, for that reason, we classify a document in an area through one of its subareas. The benefit of matching subareas is that we can provide a deeper and more detailed evaluation. The Figure 5 illustrates the subarea matching.

Document	Subarea 1	Subarea 2	...	...	Subarea N	Matched Subarea
Doc_1	0.366016	0.802408	...	...	0.566351	Subarea 2
Doc_2	0.373501	0.598308	...	...	0.665918	Subarea N
...	...	...	...	..	...	...
...	...	...	...	...	...	...
Doc_N	0.713722	0.142225	..	...	0.211479	Subarea 1

Fig. 5: Document ranking for each subarea.

If a document matches a subarea of a great area, we consider it as being part of that particular great area, for example: if a document matches Sanitation, we consider it as being part of Quality of Life. Looking at Figure 6, it can be seen that the majority of projects were semantically associated with the areas of Production and Sustainable Development. The balance of the number of projects in the areas Skilled Technologies, Quality of Life and a little less associated with Strategical shows that the university has projects that show indicators of similarity according to the strategic areas policy, established by the Brazilian Government.

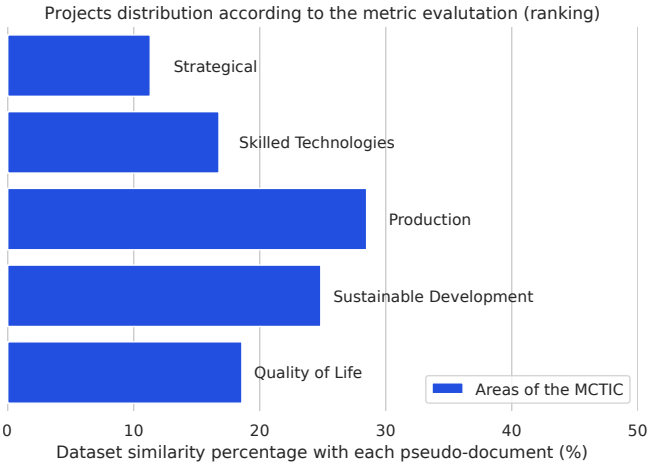

Fig. 6: Dataset ranking.

Looking further, we also investigate the distribution of projects into the most effective areas, according to the adopted methodology. In Figure 7, it is shown the subarea distribution of projects of the area Quality of Life. It can be seen in this graph that the university has a great number of projects associated with the subarea Assistive Technologies, but Sanitation needs more attention. In this case, the university may produce internal policies in order to improve the participation of researchers in this area.

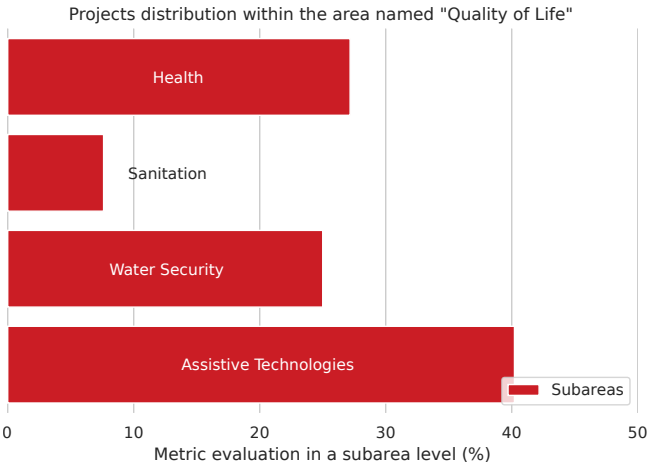

Fig. 7: Subarea level dataset ranking (Quality of Life).

Moreover, when looking at the generated distribution of the great area Skilled Technologies, shown by the Figure 8, the subarea Biotechnology carries the most of the projects, while Nanotechnology has the least number of associated projects.

VI. Final Remarks

This paper proposed a methodology using text indexing techniques, such as Latent Semantic Indexing, to automatically classify documents, composed of title and abstract of research projects from a sample of the Federal University of Tocantins and associated with the innovation policy proposed by the Brazilian Ministry of Science and Technology.

In this way, it gives almost instantly the profile of the current state of the research potentials of the university when presented a public innovation policy. It is required, though, that the policy should be organized into areas and subareas

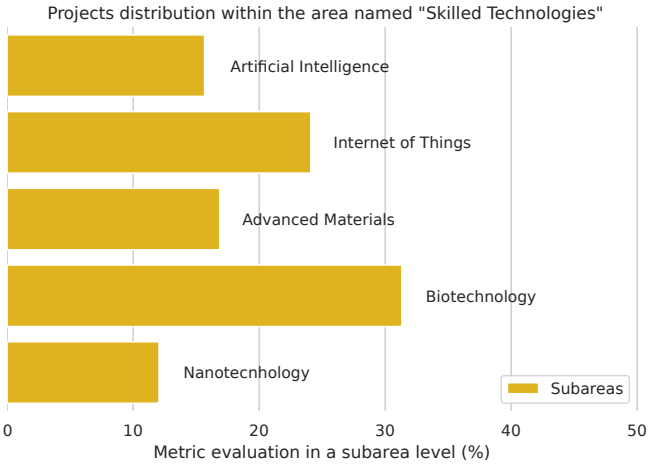

Fig. 8: Subarea level dataset ranking (Skilled Technologies).

with terms that better represent the description of such specific area. In this way, the proposed approach is capable of giving some proximity of projects into the subareas, using the LSI indexing automatically.

The proposed methodology was applied to a sample of projects from five graduate programs of the Federal University. The sample consisted of about 200 projects with title and abstract. When compared with the subareas terms, it could easily be seen that the sample shows a potential of innovation strongly centered at the great areas Production and Sustainable Development. On the other hand, the sample showed fewer projects associated with the great area Strategical. In both cases, the university can produce internal policies to either promote more projects into the strongest areas as well as to build strategies to act more effectively into the less active areas.

The main advantage of the proposed study is that it could be adapted by any research institution that wants to quickly have an overview of actuation based on the described research projects when public policies are presented by government or industry strategies. Unfortunately, this approach depends on what is written in the projects. Therefore, it is still necessary to fine-tune management to accompany the development and impacts of such projects.

As future research, it is intended to include more fields to produce a more effective way to manage the research impacts, such as to relate the projects with patents as well as papers published related to the projects.

REFERENCES

[1] C. Bacchi, *Analysing policy*. Pearson Higher Education AU, 2009, pp. X–XXI.

[2] K. Samset and T. Christensen, "Ex ante project evaluation and the complexity of early decision-making," *Public Organization Review*, vol. 17, no. 1, p. 1–17, 2015.

[3] G. Fahrenkrog, W. Polt, J. Rojo, A. Tübke, K. Zinöcker, S. A. ETH, M. Boden, S. Bührer, R. Cowan, J. Eaton *et al.*, "Rtd evaluation toolbox," *Assessing the Socio-Economic Impact of RT D-P olicies, Seville, European Commission-Joint Research Centre, IPTS*, pp. 25–27, 2002.

[4] K. Flanagan, E. Uyarra, and M. Laranja, "Reconceptualising the 'policy mix' for innovation," *Research Policy*, vol. 40, no. 5, pp. 702–713, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048733311000345

[5] MCTIC, "Portaria nº 1.122, de 19 de marÇo de 2020," 2020. [Online]. Available: https://www.in.gov.br/en/web/dou/-/portaria-n-1.122-de-19-de-marco-de-2020-249437397

[6] R. X. Coutinho, E. S. Dávila, W. M. dos Santos, J. B. Rocha, D. O. Souza, V. Folmer, and R. L. Puntel, "Brazilian scientific production in science education," *Scientometrics*, vol. 92, no. 3, pp. 697–710, 2012.

[7] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, pp. 391–407, 1990.

[8] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, pp. 26–27, 32–34.

[9] B. Rosario, "Latent semantic indexing: An overview," *Techn. rep. INFOSYS*, vol. 240, pp. 1–16, 2000.

[10] S. Sarica and J. Luo, "Stopwords in technical language processing," *Plos one*, vol. 16, no. 8, p. e0254937, 2021.

[11] A. G. Jivani *et al.*, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[12] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool, 2016, pp. 147–153, 90–108.

[13] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010, vol. 520, pp. 237–288, 451–452.

[14] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[15] S. C. Deerwester, S. T. Dumais, G. W. Furnas, R. A. Harshman, T. K. Landauer, K. E. Lochbaum, and L. A. Streeter, "Computer information retrieval using latent semantic structure," Jun. 13 1989, uS Patent 4,839,853.

[16] J. L. Bigelow, A. Edwards, and L. Edwards, "Detecting cyberbullying using latent semantic indexing," in *Proceedings of the first international workshop on computational methods for CyberSafety*, 2016, pp. 11–14.

[17] H. Chen, B. Martin, C. M. Daimon, and S. Maudsley, "Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications," *Frontiers in physiology*, vol. 4, p. 8, 2013.

[18] D. Thorleuchter and D. Van den Poel, "Improved multilevel security with latent semantic indexing," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 462–13 471, 2012.

[19] R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry, "Gene clustering by latent semantic indexing of medline abstracts," *Bioinformatics*, vol. 21, no. 1, pp. 104–115, 2005.

[20] Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized latent semantic indexing: A new approach to large-scale topic modeling," *ACM Transactions on Information Systems (TOIS)*, vol. 31, no. 1, pp. 1–44, 2013.

[21] H. Elghazel, A. Aussem, O. Gharroudi, and W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing," *Expert Systems with Applications*, vol. 57, pp. 1–11, 2016.

[22] R. V. R. D. Souza, "Análise textual dos projetos de inovação da fundação universidade federal do tocantins à luz das áreas e setores tecnológicos prioritários do ministério da ciência, tecnologia, inovação e comunicação." Master's thesis, Universidade Federal do Tocantins, Tocantins, Brazil, 2021.

[23] MCTIC, "Estratégia nacional de ciência, tecnologia e inovação: 2016-2022," 2016. [Online]. Available: http://www.finep.gov.br/images/a-finep/Politica/16_03_2018_Estrategia_Nacional_de_Ciencia_Tecnologia_e_Inovacao_2016_2022.pdf