A Dimensionality Reduction Model Applied to Documents Useful to Compliance

João Alberto da Silva Amaral
Polytechnic School of Pernambuco
University of Pernambuco UPE
Recife, Brasil
joao.amaral@cge.pe.gov.br

Prof. Dr. Fernando Buarque de Lima Neto
Polytechnic School of Pernambuco
University of Pernambuco UPE
Recife, Brasi
lfbln@ecomp.poli.br

***Abstract* — This paper proposes a semantic Natural Language Processing (NLP) approach used to assist in the automated characterization of information relevant to compliance activities. In this context, the Latent Semantic Analysis (LSA) technique was used to assist in the dimensionality reduction process. The evaluated results were achieved through the submission of two databases to the model, namely: Database of Audit reports issued by the State General Secretariat of Management (SCGE-PE - Secretaria da Controladoria-Geral do Estado, in Portuguese) of Pernambuco between the years of 2010 to 2019 and a Base of Appellate Decisions issued by the Brazilian Federal Accountability Office (TCU - Tribunal de Contas da União, in Portuguese) in 2019. The performance of two dimensionality reduction methods was evaluated: Tf-idf and LSA. To validate the results, K-means was used as a clustering technique. In addition, it was observed that the *Silhouette technique* helped us find the best cluster value for a given data sample. In the results, LSA associated with K-means presented the best performance in both databases, having achieved the best results in the TCU Base of Appellate Decisions.**

Keywords — Dimensionality reduction. Clustering. Topic modeling. Latent semantic analysis (LSA). Natural language processing. Text mining.

I. INTRODUCTION

In recent years, many companies have sought to promote an ethical and organizational culture, with the aim of reducing the incidence of fraud and financial crimes by promoting more ethical business relationships and strengthening the institution's image in the market [1]. This concern gained traction with the advent of the Sarbanes-Oxley law in the United States in 2002. Since then, several global companies have set up areas and procedures to meet the requirements of the legislation.

These areas led to the emergence of the Compliance sector in organizations, whose mission is to ensure, among other areas, the adequacy, strengthening and functioning of the Institution's Internal Control Systems, seeking to mitigate risks according to the complexity of its business, as well as disseminate the culture of controls to ensure compliance with existing laws and regulations [2]. Compliance represents the company's compliance with external and internal standards, such as laws, regulations, and corporate policies [2].

Therefore, we observed that applying a textual mining technique that assists in the execution of these tasks would be a great contribution to the compliance area, as it would bring greater efficiency, standardization and comprehensiveness in the identification and extraction of information from the analysis of documents written in natural language. Their large volumes are still a major impediment to be circumvented.

One of the biggest problems of text mining is dealing with very large-sized spaces if vector space in which each term represents a dimension is considered. Therefore, there will be as many dimensions as there are different words. There are several techniques to accomplish dimensionality reduction, LSA is a technique that combines the vector space model with mathematical scientific model of Singular Value Decomposition (SVD). This approach reveals the semantic structure underlying the text, and thus enables the retrieval of textual information a text from other texts semantically associated with it [3].

In this work the authors present a model based on topic modeling composed of data preprocessing techniques along with LSA. To validate the model, two databases were selected: Audit Reports issued by the State General Secretariat of Management and a Base of Appellate Decisions issued by the TCU.

A. Objectives

1) General Objective

The general objective of this work is to propose and evaluate an effective model, that based on unsupervised machine learning techniques, when applied to databases composed of formal documents (regulations, codes of ethics, etc.), can reduce dimensionality without loss of significant information, aiming to assist in a process of extracting information useful to compliance activity.

2) Specific Objectives

To achieve what is being described in this work, the following specific objectives were met:

- Review the literature to identify research opportunities related to the application of natural language processing in compliance;
- Implement and showcase the use of LSA as a data preprocessing technique to identify more relevant terms in the submitted texts;
- Evaluation of the performance of LSA and Tf-idf as dimensionality reduction techniques, comparing the results obtained in a clustering process using K-means.

B. Related Works

In the literature, when it comes to reducing dimensionality in texts, two strategies are still widely used currently: Zipf's Law [4], which is a technique with statistic character, says that if f is the frequency of occurrence of any word of the text, and r the ordering position according to the other words, then product $f \times r$ is approximately constant; and Luhn's proposed cut-off points [5], on which a graph $f \times$

r can a upper and lower limits defined. Words that are out of range are excluded from the analysis.

Kadhim et al. [6] propose techniques to reduce Tf-idf dimensionality and singular value decomposition (SVD) that compose a system that assists in grouping documents using the k-means algorithm. The experimental results showed that the proposed method improves the grouping performance of English text documents. The techniques used are like those used in this work, but in this case, the researchers used a configuration that selects words according to a value of relevance different from the one proposed, in addition to subjecting this model to texts in both English and Portuguese.

The work of Cai et al. [3] suggests an approach to data grouping using the traditional local K-means algorithm after projecting the set of referred textual data to a low-dimensionality space obtained by the Locality Preserving Indexing (LPI) algorithm [7] which preserves the semantic relationship between documents of the same classes.

Ding and He's proposal [8] demonstrate that Principal Component Analysis (PCA), an unsupervised dimensionality reducing technique [9] is closely linked to the also unsupervised k-means clustering method, and that both try to minimize quadratic error. Although, the authors evaluated the proposal on small textual sets, the results obtained led them to consider an evaluation of the technique in an environment distributed over new sets of data that were produced in this work, which are greater in number of objects and dimensions.

Finally, Ding and Li [10] describe an elaborate approach which combines the Linear Discriminant Analysis dimensionality reduction algorithm with the K-means algorithm, both incorporated subspace selection into a framework that has the best representation/discrimination in the analyzed data. The reduction of the dimensions in the data sets obtained by this method is achieved through the linear combination of characteristics that separate one or more groups, unlike the Latent Dirichlet Allocation algorithm, that has a probabilistic model to determine the similarity between data. The Latent Dirichlet Allocation (LDA) method [11] developed in the MapReduce programming model is explored to perform grouping and select attributes, instead of the Linear Discriminant Analysis algorithm.

II. CONCEPTUAL BASIS

A. Compliance

Compliance is perceived as the set of practices and disciplines adopted by legal entities in order to align their corporate behavior to comply with legal norms and governmental policies pertinent to the sector of operation, preventing and detecting illicit behavior, from the creation of internal structures and integrity procedures, auditing and incentives to report irregularities, which provide a diagnosis and elaborate a prognosis of conducts and their collaborators, with the effective application of codes of ethics in the respective internal scope [12].

B. Grouping or Clustering Documents

The purpose of the textual document grouping procedure is to partition a corpus collection into a given number of groups, so that similar documents are associated with the same group while unrelated documents are divided into different groups [13]. This operation occurs without the need for any prior knowledge of the submitted data.

According to Araújo Neto and Negreiros [14] the standard process of grouping documents is usually composed into six stages: 1) Pre-processing; 2) Selection of characteristics and selection of the model of document representations; 3) Selection of the dissimilarity measure; 4) Application of the grouping algorithm; 5) Cluster evaluation; and 6) Selection of descriptors for grouping.

C. Dimensionality Reduction

According to Fodor [15], the reduction of dimensionality consists in finding *d* dimensions that contain the representation of the data in a reduced manner, obtained through linear combination or not. The dimensionality of textual data sets is delimited by the number of words in the sets, so the larger the datasets and their textual body, the larger the dimensions for analysis and processing.

Due to the high dimensionality of the texts, the grouping of documents is considered as one of the most difficult tasks in the data mining area, so the use of dimensionality reduction techniques diminish complexity and ensure greater efficiency in the process of unsupervised information extraction [16].

D. LSA

Latent Semantic Analysis is a robust Algebraic-Statistical method that extracts hidden semantic structures from words and phrases, i.e., extracts characteristics that cannot be mentioned directly. These features are essential to the data, but they are not the original features of the dataset. It is an unsupervised approach, employed along with the use of Natural Language Processing (NLP).

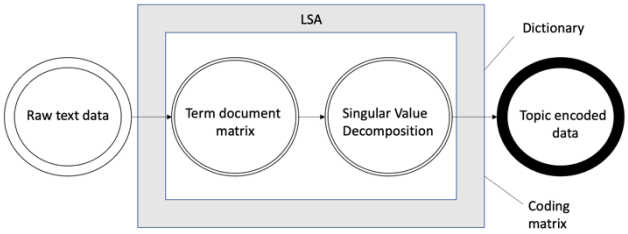

Figure 1. LSA processing. Source: Deerweste et al. [17]

The LSA algorithm consists of three main steps(see Figure 1) [17]:

1. **Input matrix creation:** The input document is represented as an array to be processed. Thus, an array of document terms is generated. Cells are used to represent the importance of each word in its respective sentence.
2. **Singular Value Decomposition (SVD)**: In this step, the decomposition of singular values is carried out in the matrix of terms of the generated document. SVD is an algebraic method that can model relationships between words, phrases and sentences. The basic idea behind SVD is that the term-document matrix can be represented as points in Euclidean space known as vectors (see Figure 2). These vectors are used to display the documents or phrases in this space. In addition to having the ability to model relationships between words and phrases, SVD can reduce noise, which helps to improve accuracy [18].

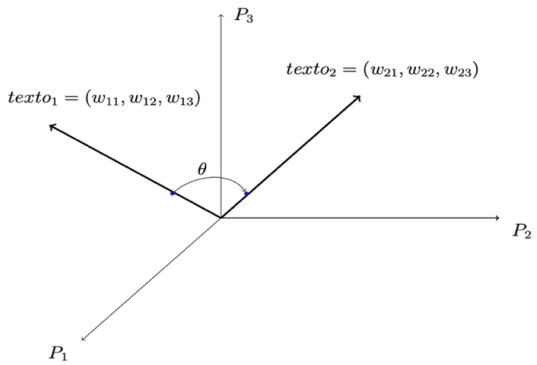

Figure 2. Term-Document Matrix. Source: Foltz [18].

3. **Selection of sentences:** Different algorithms are used to pinpoint important sentences. E.g., we use the topic method to extract concepts and sub concepts from SVD calculations.

III. METHOD

The development process was based on the CRISP-DM methodology [19], which consists of six phases: (1) Understanding the Business; (2) Understanding the Data; (3) Data Preparation; (4) Modeling; (5) Evaluation; (6) Deployment. Only the deployment step will not be covered in this paper.

A. Understanding the Business

The stage of understanding the business initially involved conducting a systematic review of the literature [20], where the identification of the main difficulties encountered by professionals working in compliance was sought. As AI is being increasingly used in this area, text mining is applied to minimize the complexity of the activity.

B. Understanding the Data

To validate the proposed model, two databases that can make up the scope of information on the work of a compliance were selected: Audit reports issued by SCGE-PE and a base of appellate decisions issued by the TCU.

Audit reports are formal documents issued by the team of auditors after the completion of a job. This report may contain several topics relevant to the knowledge of how an organ or entity is functioning, pointing out weaknesses, non-conformities, and even good practices. The base used was composed of 122 documents in PDF, DOCX and ODT formats. The SCGE-PE does not use a document management tool, so there is no standardization in formats or efficient knowledge management that facilitates the extraction of relevant information in these documents.

The base of appellate decisions from the TCU - Tribunal de Contas da União (Brazilian Federal Accountability Office, in English) are resolutions from the Plenary or the respective Chambers of the TCU on their jurisdiction and are normally used as jurisprudence by the various supervisory bodies and are therefore relevant to compliance.

C. Data Preparation

After the data collection stage, the documents were pre-processed to remove possible noise and reduce the number of attributes from the characteristic vectors from the unsupervised methods, already detailed in Section II. To achieve those results, Python language in version 3.7. was used.

Our pre-processing accompanies the following steps: 1) separate the documents into tokens, where experiments were made with unigrams, bigrams, and trigrams; 2) remove special characters (accent, numbers, and symbols) and turn all text into lowercase letters, using the NLTK[1] library; 3) tokenization; 4) removal of stopwords (words that may be considered irrelevant to the context studied); and 5) eliminate the affixes of the remaining tokens(stemming).

D. Modeling

Our modeling follows the steps suggested by A. S. Araujo et al [14]. Figure 3 presents the model used in this research, where LSA was implemented to transform each submitted document into a matrix with 100 dimensions, as according to N. Halko et al [2]. Next, ad hoc, words with absolute relevance value > 0.00004 were selected. This value was used after it was observed that the results of several experiments, where it was verified that the removal of words with absolute values lower than the chosen one did not present significant alteration in the results obtained.

E. Evaluation

To evaluate this model the solution implemented was identified in the literature as the default method for dimensionality reduction as in Kadhim et al [6]. TF-idf was used in this implementation and the first 1000 (one thousand) most frequent words were selected from the document. This choice turns all incoming documents into tuples containing the 1000 most relevant words.

In the evaluation of the results, 3 evaluation metrics were considered, being:

- **Elbow Method** – is a technique used to find the ideal number of K clusters. This method tests the variance of the data in relation to the number of clusters. The ideal K value is the one that has the lowest Within Sum of Squares (WSS) and at the same time the lowest number of clusters [21].
- **Runtime** – The time used in a complete run of an experiment.
- **Silhouette** – it is a method for interpreting and validating the consistency within data groupings. Its analysis is used to study the separation distance between resulting clusters from a given database. It is score ranges from -1.0 to 1.0, the best result being the closest to 1.

IV. RESULTS

Four experiments were executed to evaluate the results obtained, two for each selected base. Each experiment was repeated 30 times, in which the mean values and the standard

[1] "Natural Language Toolkit — NLTK 3.5 documentation." [Online]. Available: https://www.nltk.org/. [Accessed: 08-Feb-2021].

[2] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," 2010.

deviation of the evaluation metrics were calculated (Table 1 and Table 2).

In both bases, after the preprocessing steps were completed, an experiment using Tf-idf as a dimensionality reduction technique was performed. Then, in another experiment, LSA was used to compare the results.

Regarding parameterization, all algorithms followed the pattern made available in the *sklearn* library. They were, respectively: *sklearn.cluster.KMeans* [3];

The first experiments carried out were executed with the SCGE-PE audit report database. This choice was based on two reasons: knowledge of the base and the fact that it was the smallest base among the selected ones.

The experiments were set with the objective of enabling the comparison between the standard dimensionality reduction technique (Tf-idf) and LSA. Initially the experiment was run with the models with the number of *clusters (k)* ranging in a range 2<=k<=40.

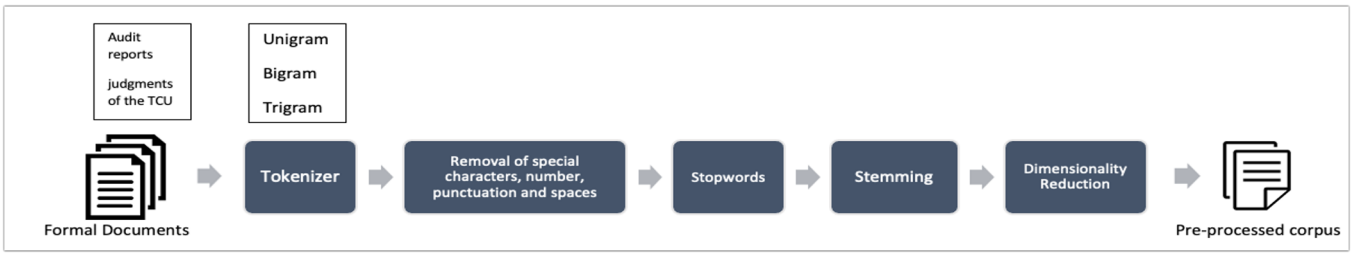

Figure 3. Suggested Model. Source: The Author

sklearn.feature_extraction.text.TfidfVectorizer[4]; and *sklearn.decomposition.TruncatedSVD*[5]. The TruncatedSVD method available in the sklearn library implements the LSA algorithm.

In all the techniques applied, the default values of the parameters were maintained, except the parameter max_features =1000, which limits the construction of the vocabulary considering only the upper max_features, ordered by term frequency throughout the corpus.

For each of the bases used in the experiments carried out in this work, a specific initial treatment was necessary:

•Audit Reports - the first base submitted was the basis of audit reports issued by SCGE-PE, the database was accessed directly in the network directories and the files were distributed in folders organized by year and area of operation. The difficulty here was to identify the final versions of each report, because in these folders there were several versions of the same document, in addition to the work papers used in the preparation of the final report. This work was carried out manually and at the end of this treatment 122 documents were selected.

•Appellate decisions from the TCU - The last base collected was the one composed of appellate decisions from the TCU in sqlite database format. The database originally contained 298,942 (two hundred and ninety-eight thousand nine hundred and forty-two) documents issued between 1992 and 30/08/2019. Due to infrastructure limitation, however, the extracted appellate decisions were only the ones issued in 2019 using the Native SQLite Manager tool and exporting the data to the CSV format. The extraction resulted in a total of 17,133 (seventeen thousand one hundred and thirty-three) judgments.

1) Scenario 1 - Audit Reports

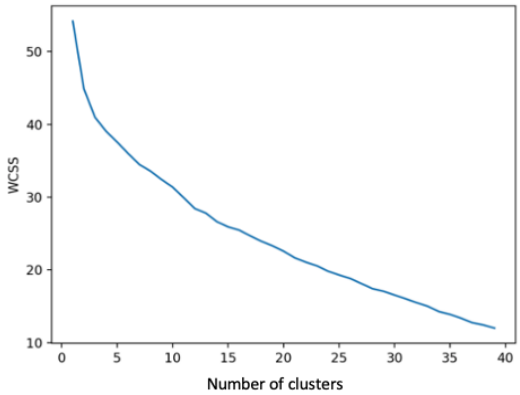

Figure 4. Elbow curve considering the database of audit reports with the Tf-idf technique. Source: The Author

The elbow method [21] was used (Figure 4 and Figure 5) to help identify the ideal number of clusters to perform the comparative analysis between the two approaches.

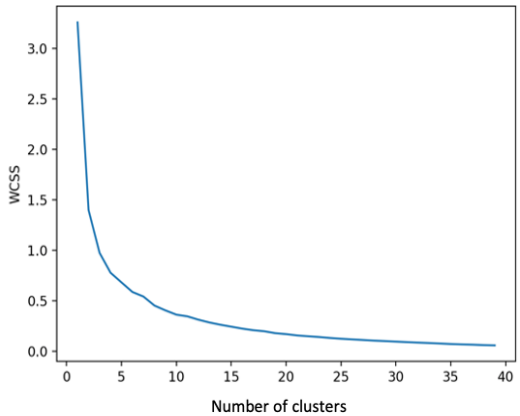

Figure 5. Elbow curve considering the database of audit reports with the LSA technique. Source: The Author

[3] "sklearn.cluster.KMeans — scikit-learn 0.24.1 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. [Accessed: 05-Feb-2021].

[4] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, 2016, pp. 61–66.

[5] "sklearn.decomposition.TruncatedSVD — scikit-learn 0.24.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html. [Accessed: 30-Apr-2021].

Observing Figure 4 and Figure 5, it is possible to identify that the best performance obtained was through modeling using the LSA technique. In Figure 4, with Tf-idf, convergence occurs in very high values of the WCSS[6] measure, what indicates poorly defined clusters. Figure 5, however, presents better convergence, noticeable through its clearer elbow curve.

Furthermore, *silhouette* (see Figure 6) was used to assess the results obtained. Here, the range of the number of clusters between 2<=k<=10 was selected, because it was verified in the analysis of elbow curves to be the best range of clusters.

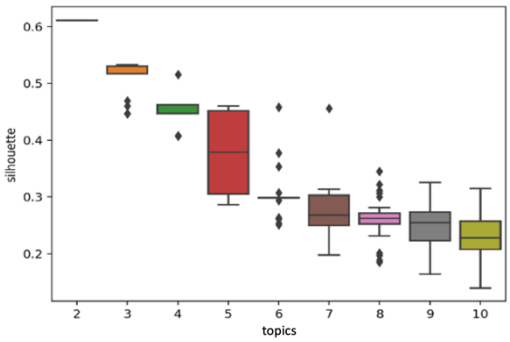

Figure 6. Boxplot of silhouette measure (Audit Reports), considering the lsa experiment approach with k-means++. Source: The Author

From the analysis of Figure 6, the value of k = 4 was utilized to compare the results obtained. This value was chosen because it presents the largest number of topics obtained without implying a large variation in the value of the silhouette. Table 1 presents the mean values and standard deviation of the four evaluation criteria used in this research.

Table 1 Evaluation of the results obtained with K-means using Tf-idf and Lsa with the basis of audit reports (experiments 01 and 02).

K-means	Dimensionality Reduction	Time	Seilhuete
k-means++	TF-IDF	0,32s (0,015)	0,125 (0,0014)
random	TF-IDF	0,25s (0,035)	0,129 (0,0009)
k-means++	LSA	0,38s (0,002)	**0,462 (0,0007)**
random	LSA	0,33s (0,012)	**0,462 (0,0003)**

Finally, the results (Table 1) of this experiment show a considerable benefit in the use of the LSA technique to the detriment of TF-idf. Considering the three-evaluation metrics used, only in one, the execution time, Tf-idf showed better results.

2) Scenario 2 - Appellate decisions from the TCU

In scenario two, all the steps followed in scenario one was repeated, now using the TCU appellate decisions as a database.

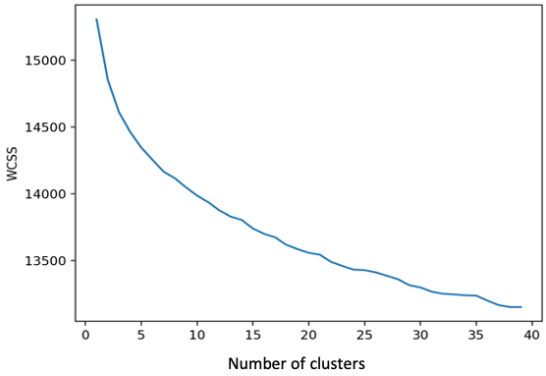

Figure 7. Elbow curve considering the TCU's database of judgments with the Tf-idf technique. Source: The Author

In Figure 7, again, there is a visible convergence occurring with the WCSS value still very high; however, the curve presents a better-defined formation than the one in scenario 1.

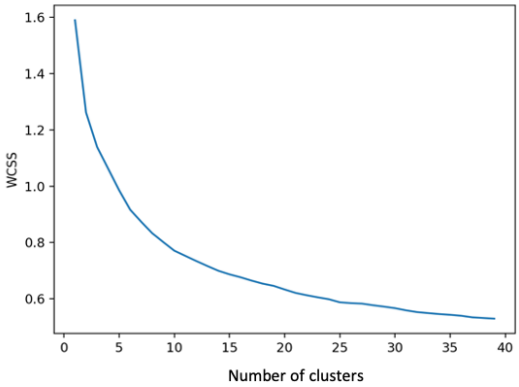

Figure 8. Elbow curve considering the database of judgments of the TCU with the LSA technique. Source: The Author

In Figure 8, once again the result resembles that obtained in scenario 1, with a better delineated elbow curve, the range 2<=k<=10 was defined as being the most promising for our evaluation sequence.

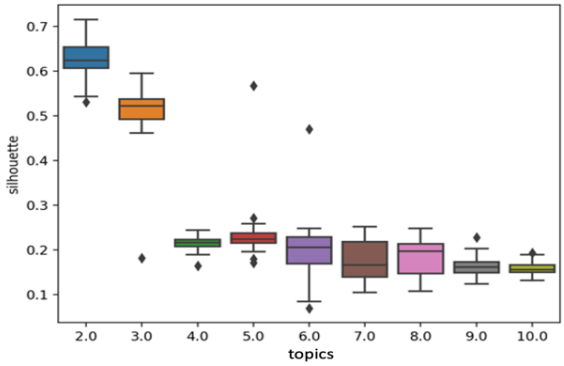

Figure 9. Boxplot of silhouette measure (TCU judgments), considering the approach of the experiment with the LSA with k-means++

In Figure 9 it is visible that the *silhouette* values for k=2 and k=3 stand out from the rest and therefore table 2 was assembled with the values obtained in the evaluation criteria. We chose k = 3 because this is the largest number of topics

[6] WCCS (Within-Cluster-Sum-of-Squares) - is the sum of squares of the distances of each data point in all clusters to their respective centroids.

obtained without implying a large variation in the value of the silhouette.

Table 2 Evaluation of the results obtained with K-means using Tf-idf and LSA based on judgments of the TCU.

K-means	Dimensionality Reduction	Time	Seilhuette
k-means++	TF-IDF	8,83s (0,032)	0,3728 (0,0127)
random	TF-IDF	5,49s (0,043)	0,317 (0,0214)
k-means++	LSA	12,68s (0,011)	**0,563 (0,0118)**
random	LSA	11,76s (0,036)	0,249 (0,0421)

Table 2 presents the mean and standard deviation values of the evaluation criteria used in this research. The results show a considerable advantage to using LSA in detriment of TF-idf in all criteria, especially the values obtained for the *silhouette.*

V. Conclusions

The main objective of this research was to conduct a study that evaluated the development of a dimensionality reduction model that helps in the extraction of information in long texts written in Portuguese, with technical or legal formalism.

As for the results obtained, the pre-processing techniques proved to be very valuable in the manipulation and qualification of the submitted documents. However, the difficulty in dealing with the audit reporting database, and the poor knowledge of the documents in the TCU judgment database, have hampered the performance of the proposed model.

Finally, based on the results presented, the use of the latent semantic analysis (LSA) shows itself to be promising in solving this problem. Furthermore, it was possible to observe that there is a gain in the quality of clusters obtained after the use of LSA.

VI. FUTURE WORK

- Submit the results of the pre-processing step to a group of experts to improve the results obtained.
- Submit our data to other dimensionality reduction techniques to identify the one that best suits our problem.
- Evolve our work to extract, in an unsupervised manner the largest number of topics that are relevant to compliance from submitted texts.

References

[1] M. de A. Coimbra and V. A. Manzi, *Manual de Compliance: Preservando a boa governança e a integridade das organizações*. São Paulo, 2010.

[2] L. LEGAL ETHICS COMPLIANCE, "Segurança da Informação e Compliance," 2012.

[3] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[4] T. J. Barnes and M. W. Wilson, "Big Data, social physics, and spatial analysis: The early years:," *http://dx.doi.org/10.1177/2053951714535365*, vol. 1, no. 1, Apr. 2014.

[5] H. P. Luhn, "A Business Intelligence System," *IBM J. Res. Dev.*, vol. 2, no. 4, pp. 314–319, Apr. 2010.

[6] A. I. Kadhim, Y. N. Cheah, and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," in *Proceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014*, 2015, pp. 69–73.

[7] Q. Gu, M. Danilevsky, Z. Li, and J. Han, "Locality Preserving Feature Learning," PMLR, Mar. 2012.

[8] C. Ding and T. Li, "Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering," 2007.

[9] I. T. Jolliffe, "Mathematical and Statistical Properties of Population Principal Components," Springer, New York, NY, 1986, pp. 8–22.

[10] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 2004, pp. 225–232.

[11] D. M. Blei, "Probabilistic topic models," in *Communications of the ACM*, 2012, vol. 55, no. 4, pp. 77–84.

[12] D. Felipe, S. Xavier, F. P. De Betim, F. P. De Betim, F. Pedro, and L. Fpl, "Compliance uma ferramenta estratégica para a segurança das informações nas organizações," in *Simpósio internacional de gestão de projetos, inovação e sustentabilidade*, 2017.

[13] A. K. Jain, "Data clustering: 50 years beyond K-means q," *Pattern Recognit. Lett.*, vol. 31, pp. 651–666, 2009.

[14] A. S. Araújo Neto and M. Negreiros, "Use of text mining techniques for unsupervised organization of digital procedural acts," *Rev. Informática Teórica e Apl.*, vol. 25, no. 4, p. 74, Nov. 2018.

[15] I. K. Fodor, "UCRL-ID-148494 A Survey of Dimension Reduction Techniques A survey of dimension reduction techniques," 2002.

[16] B. S. Santos *et al.*, "Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts," *Proc. XI Brazilian Symp. Inf. Syst. (SBSI 2015)*, no. Sbsi, pp. 667–674, 2015.

[17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[18] P. W. Foltz, "Latent semantic analysis for text-based research," *Behav. Res. Methods, Instruments, Comput.*, vol. 28, no. 2, pp. 197–202, 1996.

[19] R. Wirth and J. Hipp, "[PDF] Crisp-dm: towards a standard process modell for data mining | Semantic Scholar," in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[20] A. Pinheiro, C. Melquiades, J. Amaral, R. Cirne, and J. N. Sampaio, "Use of Artificial Intelligence Applied to Compliancetle." Em Submissão, 2020.

[21] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *Strateg. Manag. J.*, vol. 17, no. 6, pp. 441–458, Jun. 1996.