Chapter

Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches

Raphael Souza de Oliveira
and Erick Giovani Sperandio Nascimento

Abstract

The Brazilian legal system postulates the expeditious resolution of judicial proceedings. However, legal courts are working under budgetary constraints and with reduced staff. As a way to face these restrictions, artificial intelligence (AI) has been tackling many complex problems in natural language processing (NLP). This work aims to detect the degree of similarity between judicial documents that can be achieved in the inference group using unsupervised learning, by applying three NLP techniques, namely term frequency-inverse document frequency (TF-IDF), Word2Vec CBoW, and Word2Vec Skip-gram, the last two being specialized with a Brazilian language corpus. We developed a template for grouping lawsuits, which is calculated based on the cosine distance between the elements of the group to its centroid. The Ordinary Appeal was chosen as a reference file since it triggers legal proceedings to follow to the higher court and because of the existence of a relevant contingent of lawsuits awaiting judgment. After the data-processing steps, documents had their content transformed into a vector representation, using the three NLP techniques. We notice that specialized word-embedding models—like Word2Vec—present better performance, making it possible to advance in the current state of the art in the area of NLP applied to the legal sector.

Keywords: legal, natural language processing, clustering, TF-IDF, Word2Vec

1. Introduction

In recent years, the Brazilian Judiciary has been advancing toward turning all its acts digital. Following this direction, the Brazilian Labour Court implemented in 2012 the Electronic Judicial Process (acronym in Portuguese for "*Processo Judicial Eletrônico*"—PJe), and from this date, all new legal proceedings have already been born electronic. According to the Annual Analytical Report of Justice in Numbers 2020 (base year 2019) [1], produced by the National Council of Justice (acronym in Portuguese for "*Conselho Nacional de Justiça*"—CNJ), more than 99% of the ongoing cases are already on this platform.

Knowing that human beings cannot promptly analyze a large set of data, especially when such data do not appear to correlate, a way to assist in the pattern-recognition process is through statistical, computational, and data analysis methods. From the perspective that an exponential increase in textual data exists, the analysis of patterns in legal documents has become increasingly challenging.

Currently, one of the major challenges in the legal area is to respond quickly to the growing judicial demand. The Brazilian legal system provides for ways to ensure the swift handling of judicial proceedings, such as the principle of the reasonable duration of a case, the principle of speed, the procedural economy, and due process to optimize the procedural progress [2]. Therefore, with the aid of some clustering mechanism, that is, the grouping of processes, with a good rate of similarity between the documents to be analyzed, it was possible to help in the distribution of work among the advisors of the office for which the process was drawn. In addition, it contributed to the search for case law[1] for the judgment of the cases in point, to ensure a speedy trial, upholding the principle of legal certainty. According to Gomes Canotilho [3]:

> *"The general principle of legal certainty in a broad sense (thus encompassing the idea of trust protection) can be formulated as follows: the individual has the right to be able to rely on the law that his acts or public decisions involved in his rights, positions or legal relations based on existing legal norms and valid for those legal acts left by the authorities on the basis of those rules if the legal effects laid down and prescribed in the planning are connected to the legal effects laid down and prescribed in the legal order" (2003, p. 257).*

Thus, this legal management tool created positive impacts such as the decrease of the operational costs of a legal proceeding, as a result of reducing its duration, meaning lower expenses on the allocation of the necessary resources for its judgment.

Recently, machine learning algorithms have demonstrated through research that they are powerful tools capable of solving high-complexity problems using natural language processing (NLP) [4]. In this sense, it is possible to highlight the works of [5–9], which apply the techniques of word-embedding generation, a form of vector representation of terms, and consequently of documents, taking into account their context. The use of these word embeddings is essential when analyzing a set of unstructured data presented in the form of large-volume documents in court.

Nowadays, a specialist screens the documents and distributes among the team members the legal proceedings to be judged, setting up a deviation from the main activity of this specialist, which is the production of draft decisions. This contributed to an increase in the congestion rate (an indicator that measures the percentage of cases that remain pending solution at the end of the base year) and to the decrease in the meeting of demand index (acronym in Portuguese for "*Índice de Atendimento à Demanda*"—IAD—an indicator that measures the percentage of proceedings in downtime, compared to the number of new cases). It becomes evident in the consolidated data of the Labor Justice contained in **Table 1**, with data extracted from the Annual Analytical Report of Justice in Numbers 2020 (base year 2019) [1] produced by the National Council of Justice (CNJ).

This work aims, therefore, to present the degree of similarity between the judicial documents that was achieved in the inferred groups through unsupervised learning *via* the application of three techniques of NLP, namely: (i) term frequency-inverse document frequency (TF-IDF); (ii) Word2Vec with CBoW (continuous

[1] A legal term meaning a set of previous judicial decisions following the same line of understanding.

	Description	2° Degree	1° Degree	Total
Workforce				
Magistrates	Legal authority	559	3077	3636
Legal workers	Public administration employee	6911	22,785	29,696
Legal load handling				
Stockpile	Number of pending cases	792,223	3,741,548	4,533,771
New cases	Number of new cases	898,104	2,632,093	3,530,197
Judged	Number of cases judged	989,324	3,036,686	4,026,010
Closed	Number of cases with final decision	941,356	3,244,652	4,185,708
Productivity indexes				
IAD	Closed cases/new cases	104.8%	123.3%	118.6%
Congestion tax	Closed cases/(new cases + stockpile)	45.7%	53.6%	52.0%
Knowledge	Fact awareness phase	—	35.1%	35.1%
Execution	Judgment enforcement phase	—	72.7%	72.7%
Indexes per magistrate				
New cases	Average number of new cases per magistrate	1607	662	821
Workflow	Average number of cases per magistrate	3583	2794	2,927
Judged cases	Average number of cases judged per magistrate	1770	1103	1216
Closed cases	Average number of cases closed per magistrate	1684	1179	1264
Indexes per legal worker				
New cases	Average number of new cases per worker	135	83	95
Judged cases	Average number of cases judged per worker	300	351	339
Closed cases	Average number of cases closed per worker	141	148	146

Table 1.
Report of indicators of Brazilian labor justice.

bag of words) trained for general purposes for the Portuguese language in Brazil (Word2Vec CBoW pt-BR); and (iii) Word2Vec with Skip-gram trained for general purposes for the Portuguese language in Brazil (Word2Vec Skip-gram pt-BR).

This degree of congruence signals the model's performance and is set from the average similarity measure of the grouped files, based on the similarity cosine between the elements of the group to its centroid and, comparatively, by the average cosine similarity among all the documents of the group.

Aiming to delimit the scope of this research, a dataset containing information from documents of the Ordinary Appeal Interposed (acronym in Portuguese for "*Recurso Ordinário Interposto*"—ROI) type was extracted from approximately 210,000 legal proceedings. The Ordinary Appeal Interposed was used as a reference, as this is usually the type of document that induces the legal proceedings for judgment in the higher instance (2nd degree), thus instituting the Ordinary

Appeal (acronym in Portuguese for "*Recurso Ordinário*"—RO). That is a free plea, an appropriate appeal against definitive and final judgments proclaimed at first instance, seeking a review of the judicial decision drawn up by a hierarchically superior body [10].

For the present work, a literature review on unsupervised machine learning algorithms applied to the legal area was performed, using NLP, and an overview of recent techniques that use artificial intelligence (AI) algorithms in word-embedding generation. Then, we applied some methods until the results were obtained, comparing and discussing them, and finally, conclusions and future challenges were presented.

2. State-of-the-art review

Machine learning algorithms have in the most recent research demonstrated a great potential to solve high-complexity problems, which follow the categories into (i) supervised machine learning algorithms; (ii) unsupervised; (iii) semi-supervised; and (iv) by reinforcement [11]. In the context of this chapter, the literature review focused on the search for the most recent research on unsupervised machine learning or clustering algorithms applied to the legal area using NLP.

The investigation revealed that there are not many works dealing with the highlighted topic, which proves its complexity. Thus, we sought to expand the research by removing the restriction to the legal area bringing light to other publications. In [12], we discussed the content recommendation system approaches based on grouping for similar articles that used TF-IDF to perform vector transformation of the document contents and, through cosine similarity, applied k-means [13] for clustering them. In [14], the authors automatically summarized texts using TF-IDF and k-means to determine the document's textual groups used to create the abstract. Then, TF-IDF is considered the primary technique for vectorizing textual content and k-means the most used algorithm for unsupervised machine learning.

Therefore, we can assume that choosing the best technique of generating word embeddings requires investigation, experimentation, and comparison of models. Several recent pieces of research have demonstrated the feasibility of using word embeddings to improve the quality of AI algorithm results for pattern detection, classification, among other uses.

In 2013, Mikolov et al. [6] proposed two new architectures to calculate vector representations of words calling them Word2Vec, which was considered, at the time, as a reference in the subject. Subsequently, techniques of word embeddings based on the use of the long short-term memory network (LSTM) [15] became widely used for speech recognition, language modeling, sentiment analysis, and text prediction, and that, unlike the recurrent neural network (RNN) they can forget, remember and update the information thus taking a step forward from the RNNs [16]. Therefore, LSTM-based libraries, such as Embeddings from Language Models (Elmo) [17], Flair [18], and context2vec [19] created a different word embedding for each occurrence of the word, related to the context, that allowed to capture the meaning of the word.

In more recent years, new techniques of word embeddings have emerged, with emphasis on (i) Bidirectional Encoder Representations from Transformers (BERT) [9], context-sensitive model with architecture based on a transformer model [20]; (ii) Sentence BERT (SBERT) [21], a "Siamese" BERT model that was proposed to improve BERT's performance when seeking to obtain the similarity of sentences; and (iii) Text-to-Text Transfer Transformer (T5) [22], a framework for treating NLP issues as a text-to-text problem, that is, input to the template as text and template output as text.

From this analysis, it was possible to advance in the current state of the art in the area of NLP applied to the legal sector, by conducting a comparative study and application of the techniques TF-IDF, Word2Vec CBoW, and Word2Vec Skip-gram to perform the grouping of labor legal processes in Brazil using the k-means algorithm and the cosine similarity.

3. Methodology

This section presents each step necessary to achieve the results and to make it possible to analyze them comparatively. To perform all the implementations of the routines necessary for this study, the Python programming language (version 3.6.9) was used and, among other libraries, (i) Numpy (version 1.19.2) was used; (ii) Pandas (version 1.1.3); (iii) Sklearn (version 0.21.3); (iv) Spacy (version 2.3.2); and (v) Nltk (version 3.5).

Every processing flow (pipeline) consists of the phases: (i) data extraction; (ii) data cleansing; (iii) generation of word-embedding templates; (iv) calculation of the vector representation of the document; (v) unsupervised learning; and (vi) calculation of the similarity measure, as detailed in the following subsections.

3.1 Data extraction

The dataset used for these studies belongs to the Regional Labour Court of the 5th Region (acronym in Portuguese for "Tribunal Regional do Trabalho da 5ª Região"—TRT5). There are approximately 210 (two hundred and ten) thousand documents of the Ordinary Appeal Interposed type, incorporated into the Electronic Judicial Process (PJe) system, originally added to the PJe in portable document format (PDF) or hypertext markup language (HTML). As the PJe has a tool for extracting and storing the contents of documents, there was no need for further processing in obtaining the text of such files.

In addition to the content of the documents, the following information was extracted: (i) the name of the parts of the proceedings to which such documents belonged; (ii) the list of labor justice issues from the Unified Procedural Table[2] (acronym in Portuguese for "*Tabela Processual Unificada*"—TPU) of the Labour Justice branch (made available by the National Council of Justice [CNJ] and consolidated by the Superior Labour Court [acronym in Portuguese for "*Tribunal Superior do Trabalho*"—TST]); and (iii) list of abbreviations (acronyms) with their full translation according to tables made available by the Supreme Court (acronym in Portuguese for "Supremo Tribunal Federal"—STF).[3]

3.2 Data cleaning

Preprocessing is a fundamental step for the application of artificial intelligence techniques and involves the following: (i) data standardization (when there is a large discrepancy between the values presented to the technique); (ii) the withdrawal of null values; and (iii) the reorganization and adequacy of the structure of the dataset. In this case, it is usually necessary for experts to conduct an exploratory analysis of the data used in advance to determine the direction of preprocessing.

[2] Labour Justice Unified Procedural Table. Available at: https://www.tst.jus.br/web/corregedoria/tabelas-processuais

[3] Table of abbreviations (and acronyms) made available by the Supreme Court. Available at: https://www.stf.jus.br/arquivo/cms/publicacaoLegislacaoAnotada/anexo/siglas_cf.pdf

For this phase, this study uses two forms of preprocessing: (i) detection of the subjects of the Unified Procedural Table (contained in the extracted documents) and (ii) cleaning the contents of the documents.

For the detection of the subjects of the TPU present in the extracted documents, regular expression matching was used as the search technique to measure the occurrences of these words in the files marking them with "tags" referring to the subject found.

For cleaning the contents of documents, usually using a regular expression, the steps were as follows:

- HTML tags: removed the html tags found in the document, such as <script>, <body>, <style> etc.;
- TPU subjects: replaced the subject text with a subject tag, for example, "*hora extra*" (overtime) changed to *hora_extra*;
- Related Persons: replaced the name of the individuals linked to the legal cases of the documents, such as the name of the author(s) and defendant(s), by the "tag" "*parteprocesso*" (part in the process);
- Judicial process number: replaced the number of the judicial process (according to the standard formatting defined nationally by the CNJ, NNNNNNN-NN. NNNN.N.NN.NNNN where N is a numeral) by the "tag" "*numeroprocesso*" (process number);
- Standardization of abbreviations: replacement of abbreviations (acronyms) by the full translation as drawn STF list as reported in Section 3.1, for example, CLT was transformed into "*Consolidação das Leis do Trabalho*" (Consolidated Labour Law);
- Addresses: replaced the addresses contained in the document with the "tag" "*enderecoprocesso*" (addresses in the process);
- Links: removed Internet links contained in the text;
- Date and Time: replacement of date and time content with "*datahora*" (datetime) tag;
- Time: replacement of the time content with the "*hora*" (hour) tag;
- Days of the week: removed the days of the week found in the document;
- Document ids: replacement of PJe document ids referenced in the document with "tag" "*sequenciadocumento*" (document sequence). These ids are typically composed of alphanumeric characters;
- Unit of measure: replaced the units of measurements and their values by the "tag" "*unidademedida*" (measurement unit);
- Numbers: replaced the numbers in full, ordinal numbers, and numerical sequences by the "tag" "*numeral*" (number);
- Judging bodies: replaced the judging bodies (e.g., "*Tribunal Regional do Traabalho*" [Regional Labour Court]) by the "*orgaojulgador*" (organjudge) tag;

- Months of the year: removed the months of the year found in the document;
- Judicial Stopwords: only when the technique employed is TF-IDF. The common words were removed in all texts of the judiciary, such as (i) "*magistrado*" (magistrate) and (ii) "*processo*" (legal proceeding), among others;
- Stopwords:
  - TF-IDF: removed all stopwords from the Portuguese language, such as "*de*" (from), "*da*" (of), "*a*" (the), "*o*" (the), "*esta*" (this) etc.;
  - Other techniques: removed only the non-adverbs of the Portuguese language, for example, the words "*não*" (no), "*mais*" (more), "*quando*" (when), "*muito*" (very), "*também*" (also), and "*depois*" (after) remain in the document;
- Line breaks: replaced line breaks by space;
- Punctuation marks:
  - TF-IDF: removed all the punctuation marks contained in the documents;
  - Other techniques: removed the punctuation marks except dot (.), comma (,), exclamation (!), and interrogation (?);
- Lemmatization:
  - TF-IDF: applied the technique to replace words with its root, for example, words such as "*tenho*" (have), "*tinha*" (had), and "*tem*" (have) had belong of the same root "*ter*" (have);
  - Other techniques: lemmatization has not been applied;

In addition to the preprocessing detailed above, when the technique used was TF-IDF, the tags inserted in the text during this phase were removed.

3.3 Generation of word-embedding templates

An essential technique in solving machine learning problems, involving NLP, is the use of vector representation of words, in which numerical values indicate some correlation of words in the text. This chapter uses word embeddings generated and shared for the Portuguese language, such as Word2Vec CBoW and Word2Vec template with Skip-gram. These templates were created based on more than 1 billion and 300,000 tokens, with results published in the article "Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks" presented at the Symposium in Information and Human Language Technology - STIL 2017 [23].

3.4 Calculation of the vector representation of the document

Different from the TF-IDF technique, which has the vector representation of the document based on the statistical measurement of each term of the document in relation to all known corpus, and whose vector dimension is equal to the size of

the vocabulary of the corpus, the other techniques (i) Word2Vec CBoW ptBR and (ii) Word2Vec Skip-gram pt-BR need to go through a change to calculate the vector representation of the document (document embeddings). This happens because for these techniques what you can get is the vector representation of the word (word embeddings).

Thus, to calculate the vector representation for the documents some alternatives are suggested, such as (i) average of the word embeddings of the words of the document; (ii) sum of the word embeddings of the words in the document by pondering them with the TF-IDF and then dividing by the sum of the TF-IDF of the words of the document; and (iii) weighted average with the TF-IDF of the word embeddings of the words of the document, the latter being the technique chosen for presenting the best result.

3.5 Unsupervised learning

The use of unsupervised learning techniques is relevant when the intention is to detect patterns among court documents. The k-means algorithm, whose basic concepts were proposed by MacQueen [13], is the technique adopted in this study. In general, this technique seeks to recognize patterns from the random choice of K initial focal points (centroid), where K is the number of groups that one wishes to obtain and, iteratively, position the elements whose Euclidean distance is the minimum possible concerning the centroid of the group.

Since one does not have an ideal K to offer the algorithm, an approach usually used to support such a decision is to calculate the inertia, based on how well the dataset was grouped through k-means.

The inertia calculation is based on the sum of the square of the Euclidean distance from each point to its centroid and seeks to obtain the lowest K with the lowest inertia. However, the higher the K value reaches, the tendency is that inertia will be lower, and then, the elbow method was used to find the point where the reduction in inertia begins to decrease.

Hence, 31 values for K were used within the range from 30 to 61, considering an interval for each unit, selecting the K that generated the best grouping. In addition, the strategy of creating submodels, limited to two, was used for the documents of the groups whose average similarity rate did not reach a value greater than 0.5.

3.6 Similarity measure calculation

The similarity measure is an important tool for the measurement of the quality of inferred groups. In this study, the cosine similarity measure is adopted, which is a measure that calculates the cosine of the angle between two vectors projected in the multidimensional plane, the result of which is between 0 and 1, in which 1 represents that the two vectors are totally similar, and 0 represents that they are totally different. Given two vectors, X and Y, the cosine similarity is presented using a scalar product according to Eq. (1).

$$\text{similarity} = \cos(\theta) = \frac{X \cdot Y}{|X| \cdot |Y|} \tag{1}$$

Consequently, to decide whether, after the clustering of the chief model, it was necessary to generate up to two more submodels, using the average cosine similarity among all elements of the group. Although the computational cost of calculating

the similarity between all files in the group is relevant, we sought to reduce the distance between documents that were part of the same group, although they were located near the centroid. To assess the final efficiency of the technique, another form of calculation was adopted, computing for each group the average cosine similarity between the group elements and its centroid. Thus, as a measure of global similarity of each approach, we calculated the average of the average of the groups, so that the one that reached a value closer to 1 (one) was considered the best technique.

4. Results and discussions

This research shows, as per the methodology presented in the previous sections, how machine learning algorithms associated with NLP techniques are important allies in optimizing the operational costs of the judicial process. It is evidenced from the result, for example, of document screenings and procedural distribution, which allows an expert to devote oneself to their chief activity optimizing working time.

While using the k-means unsupervised learning algorithm, it was necessary to choose the best K for each NLP technique studied. In this scenario, the elbow method was applied based on the calculated inertia of each of the 31 K tested, as shown in **Figure 1**, thus achieving a better result for each technique.

From the attainment of the best K, the k-means model was trained and, from the grouping performed by this technique, we could reach the average similarity between the documents of each group. Those groups that did not make the cutting line of at least 0.5 of average had the group files submitted for creating up to two

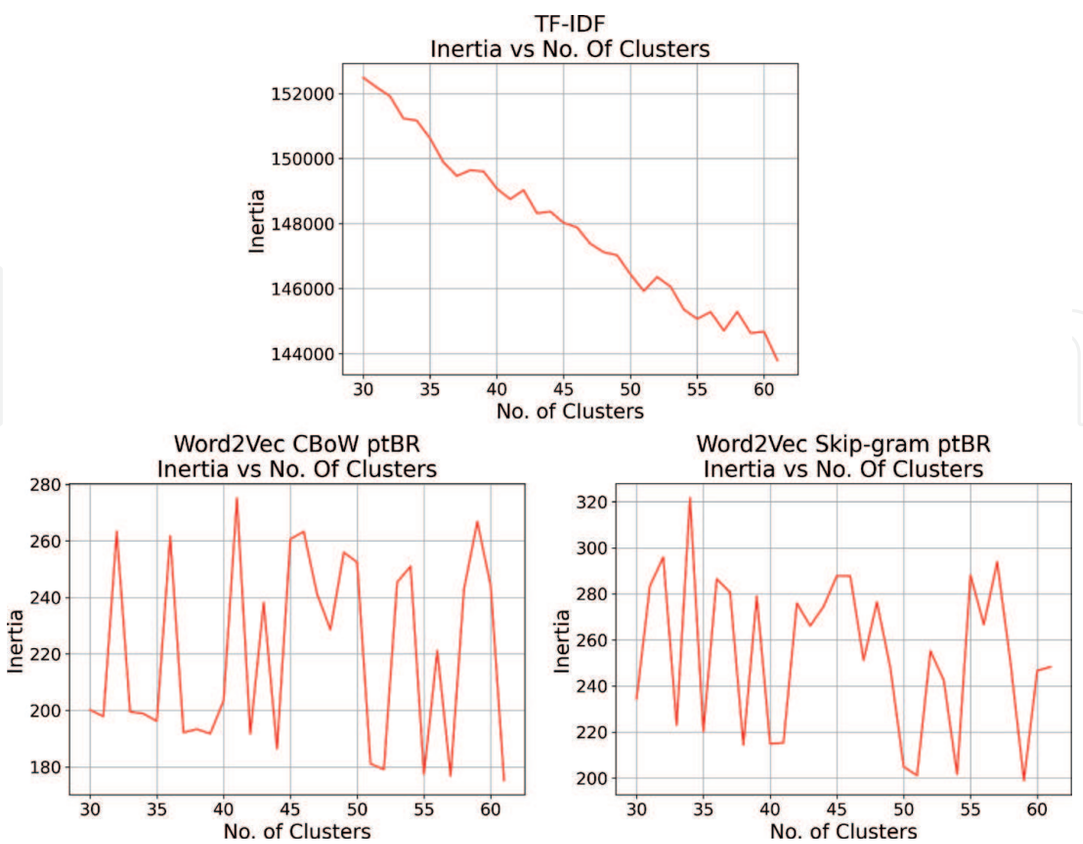

Figure 1.
Inertia charts constructed by using the elbow method for determining the best number of clusters for each approach.

submodels. As expected, only for TF-IDF technique groupings is there a need to generate submodels to improve performance.

Table 2 shows the average similarity of the groups obtained using the TF-IDF technique, as well as the result of the Word2Vec CBoW pt-BR technique. It achieved a little better measure of similarity than the Word2Vec Skip-gram pt-BR technique; however, the latter achieved its result with a smaller number of groups, which places it, in general, as the best technique.

After the groups were formed, the statistical data resulting from each approach were calculated, as shown in **Table 3** and in the comparative graph of distributions between the techniques (**Figure 2**). The cosine similarity of the group elements to

	Model		Submodel 1		Submodel 2		Final	
Type	**Groups**	**Mean**	**Groups**	**Mean**	**Groups**	**Mean**	**Groups**	**Mean**
TF-IDF	37	0.3696	43	0.4001	48	0.4002	48	0.4002
Word2Vec CBoW ptBR	59	0.9060	—		—		59	0.9060
Word2Vec Skip-gram ptBR	34	0.9044	—		—		**34**	**0.9044**

Table 2.
Mean cosine similarity between all elements of the group. The best results are highlighted in bold.

Type	**Groups**	**Mean**	**Std.**	**Min.**	**25%**	**50%**	**75%**	**Max.**
TF-IDF	49	0.6241	0.1718	0.2466	0.5021	0.5864	0.1639	0.9644
Word2Vec CBoW ptBR	59	0.9475	0.0632	0.7640	0.9352	0.9790	0.991	0.9999
Word2Vec Skip-gram ptBR	**34**	**0.9481**	**0.0609**	**0.7960**	**0.9248**	**0.9763**	**0.9924**	**0.9995**

Table 3.
Statistics of the cosine similarity of the group elements to the centroids. The best results are highlighted in bold.

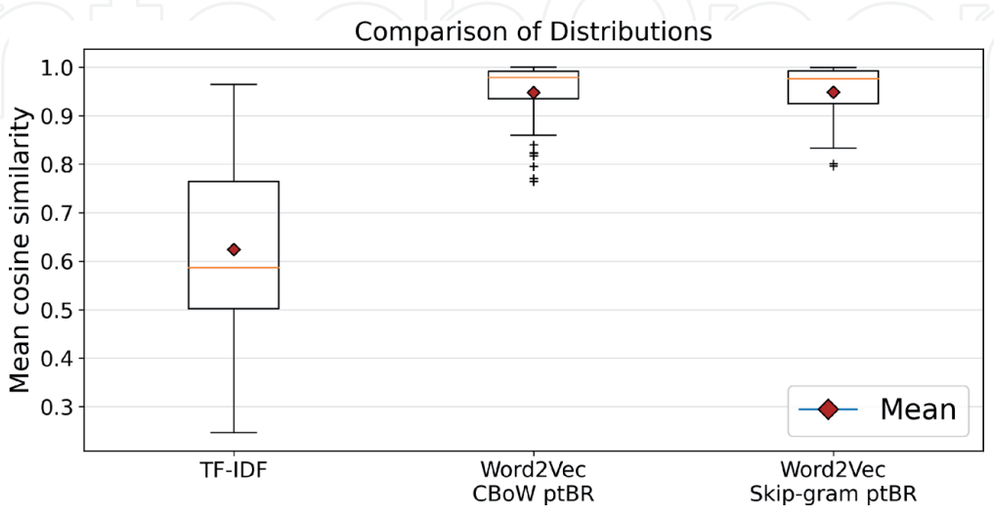

Figure 2.
Boxplots showing the distributions of the clusters calculated by each technique. The more cohesive the boxes and the less number of outliers, the better.

its centroid was used as a metric, showing the proximity of the results between the techniques with Word2Vec and highlighting the technique Word2Vec Skip-gram ptBR for the smaller amount of generated groups.

When comparing the values presented in **Tables 2** and **3**, it is noteworthy that the results presented in **Table 2** are worse in all cases. It is inferable from this observation that the similarity measure calculations shown in **Table 2** can reduce the similarity rates since there may be elements in the group positioned on completely opposite sides. From **Figure 2**, it is also possible to verify that the groupings generated by the Word2Vec technique were more cohesive than those generated by the TF-IDF technique, especially the Word2Vec Skip-gram technique, which created fewer groupings in the range of outliers than Word2Vec CBoW, demonstrating its superiority by allowing fewer groups but maintaining consistent quality and cohesion.

Given the aforesaid, among all the techniques evaluated, the Word2Vec Skip-gram pt-BR technique presented itself as the best option for word embeddings for clustering legal documents of the Ordinary Appeal Interposed type. Although the Word2Vec CBoW pt-BR technique achieves slightly better rates, it stands out from the previous one for reaching a much smaller number of groups.

The result achieved by each approach can be visualized by projecting in two dimensions of the groups formed from the three techniques: (i) TF-IDF; (ii) Word2Vec CBoW pt-BR; and (iii) Word2Vec Skip-gram pt-BR, respectively, presented in **Figures 3-5**. It is evident in the figures that the groups formed from Word2Vec are much better defined, especially skip-gram, which confirms the findings previously explained in this work.

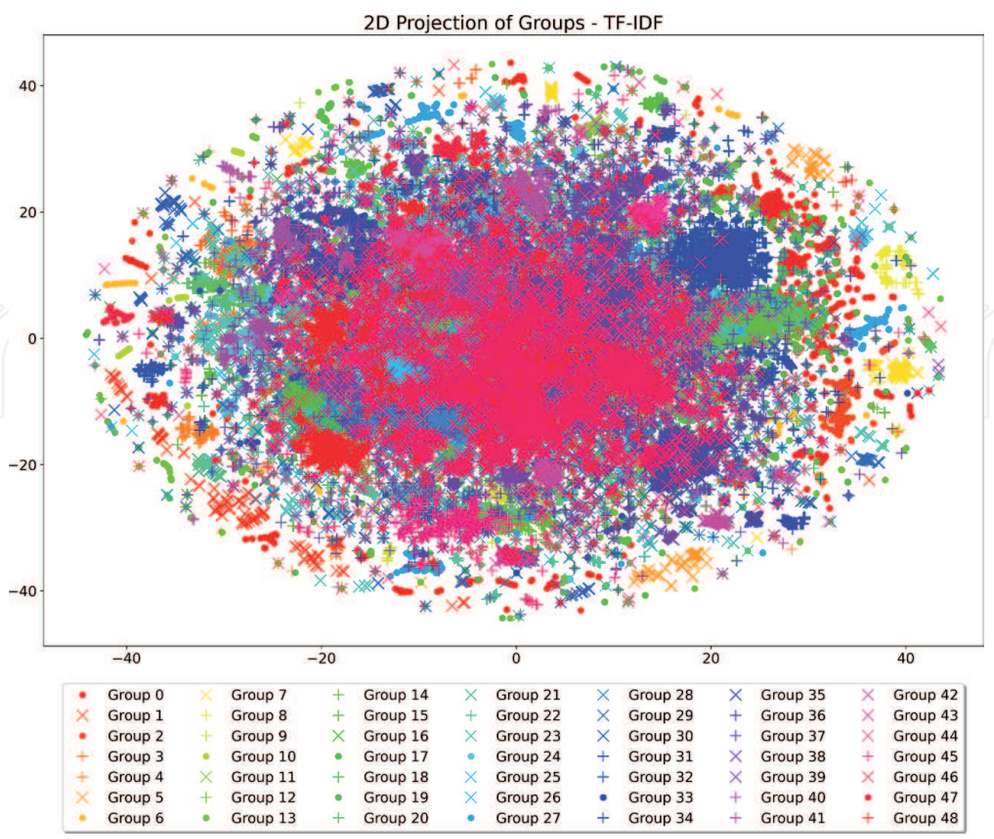

Figure 3.
2D projection of the entire test dataset, showing for each document its corresponding group formed by TF-IDF.

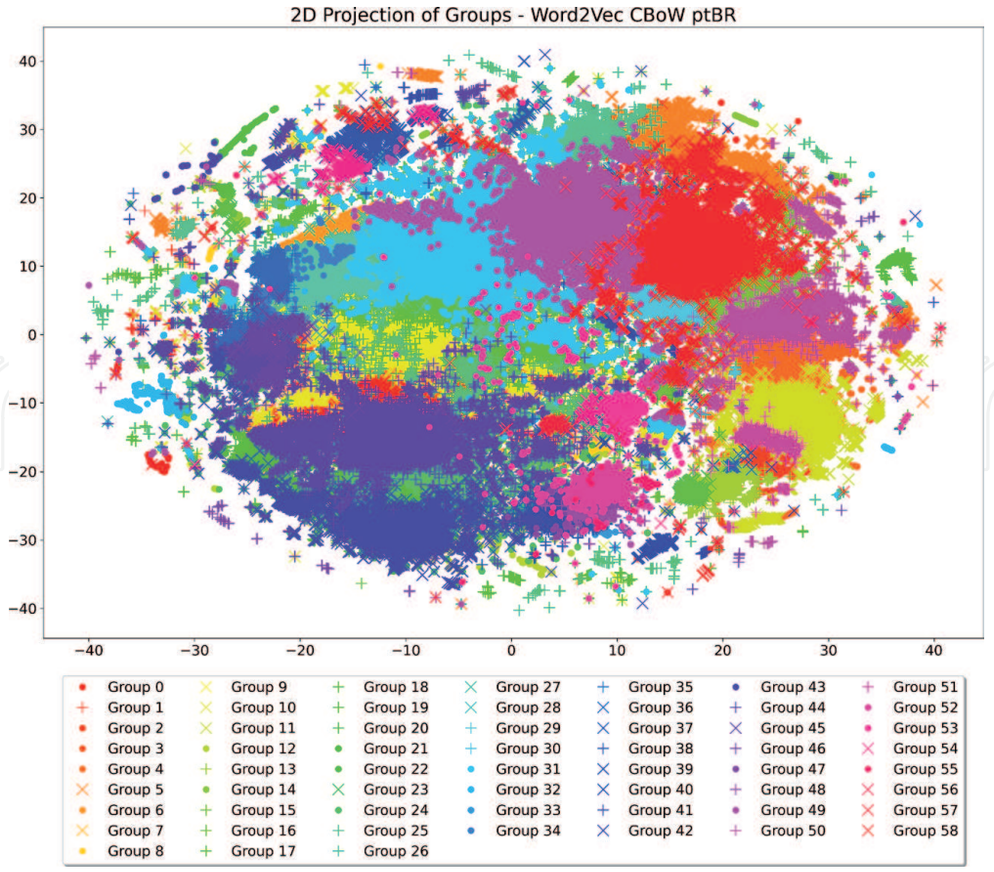

Figure 4.
2D projection of the entire test dataset, showing for each document its corresponding group formed by Word2Vec CBoW ptBR.

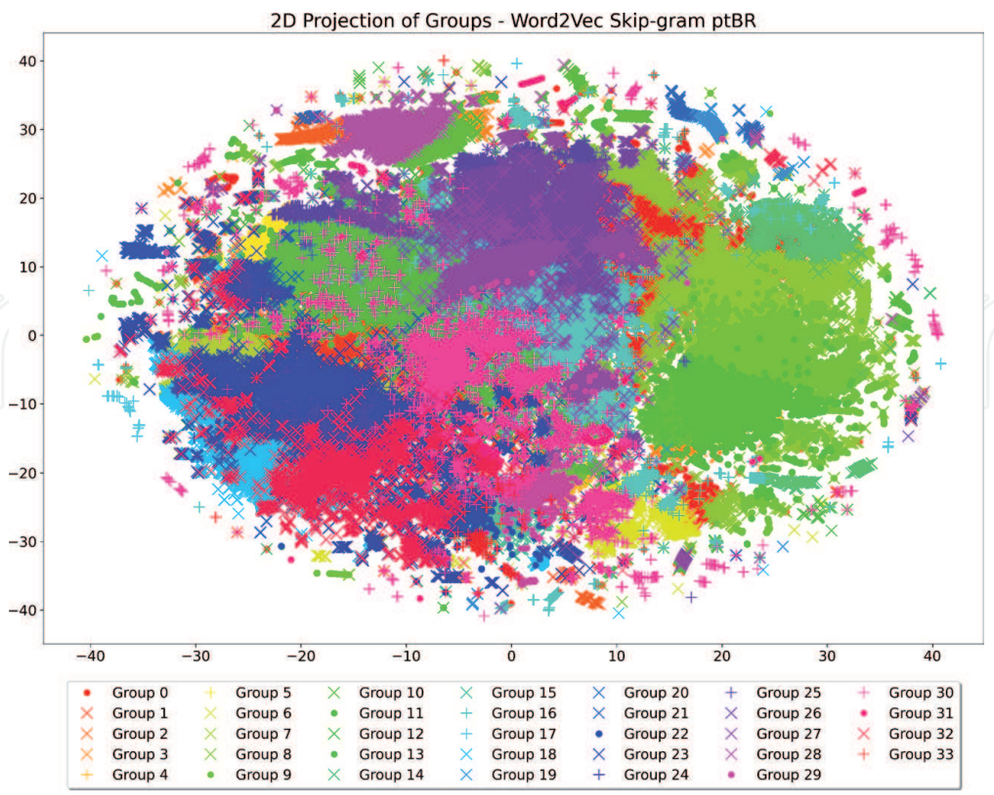

Figure 5.
2D projection of the entire test dataset, showing for each document its corresponding group formed by Word2Vec skip-gram ptBR.

5. Conclusion and future work

The use of AI as a standard detection tool based on documents from the judiciary has generally proved to be a viable and helpful solution in the scientific, technological, and practice of legal work. In this chapter, it was possible to present the results considered very promising due to the improvement in the average similarity rate. Thus, we demonstrate the possibility of using word-embedding generation techniques applied on clustering of Ordinary Appeal Interposed using AI algorithms.

Of all the techniques evaluated, the Word2Vec Skip-gram pt-BR technique presented itself as the best option for word embeddings for clustering legal documents of the Ordinary Appeal Interposed type.

We believe that specialized word embeddings have great potential in improving the results. Therefore, comes the suggestion for future study of Word2Vec specialized for the judiciary, in addition to evaluating whether the new embeddings generated provide an opportunity to improve the overall performance of clustering. In addition, using transformer-based techniques, such as BERT, can achieve promising results, using both the Portuguese language word-embedding model and training a specialized BERT model for the judiciary.

Moreover, new possibilities arise for using the techniques discussed in this chapter, such as the draft generation of decisions and classification of documents and processes.

Acknowledgements

The authors thank the Regional Labour Court of the 5th Region for making datasets available to the scientific community and contributing to research and technological development. The authors also thank the Artificial Intelligence Reference Centre and the Supercomputing Centre for Industrial Innovation, both from SENAI CIMATEC.

Author details

Raphael Souza de Oliveira[1] and Erick Giovani Sperandio Nascimento[2*]

1 TRT5—Regional Labor Court of the 5th Region, Salvador, BA, Brazil

2 SENAI CIMATEC—Manufacturing and Technology Integrated Campus, Salvador, BA, Brazil

*Address all correspondence to: ericksperandio@gmail.com

References

[1] CNJ—Conselho Nacional de Justiça. Relatório Analítico Anual da Justiça em Números 2020. 2020. Available from: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/ [Accessed: June 07, 2021]

[2] da Costa Salum G. A duração dos processos no judiciário: aplicação dos princípios inerentes e sua eficácia no processo judicial [Internet], Âmbito Jurídico, Rio Grande. Vol. XIX(145). 2016. Avaliable from: https://ambitojuridico.com.br/cadernos/direito-processual-civil/a-duracao-dos-processos-no-judiciario-aplicacao-dos-principios-inerentes-e-sua-eficacia-no-processo-judicial/ [Accessed: September 01, 2021]

[3] Canotilho JJG. Direito constitucional e teoria da constituição. 7th ed. Coimbra: Almedina; 2003

[4] Khan W, Daud A, Nasir J, Amjad T. A survey on machine learning models for Natural Language Processing (NLP). Computer Science and Engineering. 2016;**43**:95-113

[5] Wang Y, Cui L, Zhang Y. Using Dynamic Embeddings to Improve Static Embeddings. In: arXiv Preprint. arXiv:1911.02929v1. 2019

[6] Mikolov, T, Chen, K, Corrado, G, Dean, J. Efficient Estimation of Word Representations in Vector Space. In: ICLR: Proceeding of the International Conference on Learning Representations Workshop Track, Arizona, USA. 2013.

[7] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. pp. 1532-1543

[8] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;**5**:135-146

[9] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics. 2019; 1:4171-4186. DOI: 10.18653/v1/N19-1423

[10] Oliveira FJV. Os recursos na Justiça do Trabalho [Internet]. Available from: http://www.conteudojuridico.com.br/consulta/Artigos/24853/os-recursos-na-justica-do-trabalho [Accessed: June 10, 2021]

[11] Sil R, Roy A, Bhushan B, Mazumdar AK. Artificial Intelligence and Machine Learning based Legal Application: The State-of-the-Art and Future Research Trends. In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS); 18-19 October 2019; Greater Noida, India: IEEE; 2019. p. 57-62. DOI: 10.1109/ICCCIS48478.2019.8974479

[12] Renuka S, Raj Kiran GSS, Rohit P. An unsupervised content-based article recommendation system using natural language processing. In: Jeena Jacob I, Kolandapalayam Shanmugam S, Piramuthu S, Falkowski-Gilski P, editors. Data Intelligence and Cognitive Informatics (Algorithms for Intelligent Systems). Singapore: Springer; 2021. pp. 165-180. DOI: 10.1007/978-981-15-8530-2_13

[13] MacQueen J. Some methods for classification and analysis of

multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; Berkeley, CA: University of California Press; Vol. 1. 1967. pp. 281-297.

[14] D'Silva J, Sharma U. Unsupervised automatic text summarization of Konkani texts using K-means with Elbow method. International Journal of Engineering Research and Technology. 2020;**13**:2380. DOI: 10.37624/IJERT/13.9.2020.2380-2384

[15] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;**9**:1735-1780. DOI: 10.1162/neco.1997.9.8.1735

[16] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena. 2020;**404**:132306

[17] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. pp. 2227-2237. DOI: 10.18653/v1/N18-1202

[18] Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. pp. 1638-1649

[19] Melamud O, Goldberger J, Dagan I. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning; Berlin, Germany: Association for Computational Linguistics; 2016;. p. 51-61. DOI: 10.18653/v1/K16-1006

[20] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000-10. (NIPS'17).

[21] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019 . p. 3982-92. DOI: 10.18653/v1/D19-1410

[22] Roberts A, Raffel C, Lee K, Matena M, Shazeer N, Liu PJ, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In: arXiv Preprint. arXiv:1910.10683. 2019

[23] Hartmann NS, Fonseca ER, Shulby CD, Treviso MV, Rodrigues JS, Aluísio SM. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: Proceedings of the 11th Brazilian Symposium on Information and Human Language Technology (STIL). Uberlândia, Minas Gerais, Brazil: Brazilian Computing Society - SBC; 2017. p. 122-31.