Impact of Text Specificity and Size on Word Embeddings Performance: An Empirical Evaluation in Brazilian Legal Domain

Thiago Raulino Dal Pont[1(✉)], Isabela Cristina Sabo[2], Jomi Fred Hübner[1], and Aires José Rover[2]

[1] Department of Automation and Systems, Federal University of Santa Catarina, Florianópolis, SC 88040-900, Brazil
thiagordalpont@gmail.com, jomi@das.ufsc.br
[2] Department of Law, Federal University of Santa Catarina, Florianópolis, SC 88040-900, Brazil
isabelasabo@gmail.com, aires.rover@ufsc.br

Abstract. Word embeddings is a text representation technique capable of capturing syntactic and semantic linguistic patterns and of representing each word as an n-dimensional dense vector. In the domain of legal texts, there are trained word embeddings in languages like English, Polish, and Chinese. However, to the best of our knowledge, there are no embeddings based on Portuguese (Brazilian and European) legal texts. Given that, our research question is: does the specificity and size of the text corpus used for a word embedding training contribute to a more successful classification? To answer the question, we train word embeddings models in the legal domain with different levels of specificity and size. Then we evaluate their impact on text classification. To deal with the different levels of specificity, we collect text documents from different courts of the Brazilian Judiciary, in hierarchical order. We used these text corpora to train a word embeddings model (GloVe) and then had then evaluated while classifying processes with a deep learning model (CNN). In a context perspective, the results show that in word embeddings trained on smaller corpora sizes, text specificity has a higher impact than for large sizes. Also, in a corpus size perspective, the results demonstrate that the greater the corpus size in embeddings training, the better are the results. However, this impact decreases as the corpus size increases until a point where more words in the corpus have little impact on the results.

Keywords: Word embeddings · Legal corpora · GloVe · Text classification · Convolutional Neural Network

T. R. Dal Pont and I. C. Sabo—This research was supported by grants from CNPq (National Council for Scientific and Technological Development) and CAPES (Coordination for the Improvement of Higher Education Personne).

© Springer Nature Switzerland AG 2020
R. Cerri and R. C. Prati (Eds.): BRACIS 2020, LNAI 12319, pp. 521–535, 2020.
https://doi.org/10.1007/978-3-030-61377-8_36

1 Introduction

Text classification is an important part of Text Mining (TM) and Natural Language Processing (NLP) and it has been applied in many contexts [9,19,27]. Since texts are unstructured data, we can to transform them into a structured format so that it is possible to perform supervised learning using an available classifier, such as Support Vector Machines (SVM) or Convolutional Neural Networks (CNN) [18].

One of the many methods to represent text in a structured format is the Vector Space Model (VSM), where each document is represented through a numerical vector. This representation can be created using the Bag of Words (BOW) model where the vector values may be frequencies of each word in the text, generating sparse and high-dimensional representations [4]. New models of VSM representations have been proposed recently, such as Word2Vec [22], GloVe [24], and FastText [7]. They use a technique known as word embeddings. It represents each word as an n-dimensional dense vector, capable of capturing syntactic and semantic linguistic patterns [36]. To learn word embeddings, an unsupervised technique, such as GloVe, is applied to a large text corpus. Its size and the covered subjects have an impact on the quality of the embeddings and its performance in text classification. Commonly, word embeddings models use texts from several contexts. However, embeddings trained using only texts related to the classification task may get better results [20].

In the domain of legal texts, there are trained word embeddings in languages like English [10] and Polish [29]. However, to the best of our knowledge, there are no embeddings based on Portuguese legal texts. We have searched the main knowledge bases (Scopus, IEEE Xplore, ACM DL and Web of Science) for papers published in the last ten years. Nevertheless, Portuguese embeddings in multi-genre texts were already investigated [15,25].

Therefore, our research question is: does the specificity and size of the text corpus used for a word embedding training contribute to a more successful classification? To answer the question, we train word embedding models in the legal domain with different levels of specificity and size. Then, we evaluate their impact on classification. To deal with the different levels of specificity, we collect texts from different courts of the Brazilian Judiciary, in hierarchical order. These text corpora are used to train a word embeddings model (GloVe) and then are evaluated while classifying processes with a deep learning model (CNN).

The motivation for our research is the lack of academic work about word embeddings applied to legal texts in Portuguese (Brazilian and European). Moreover, few papers evaluate the influence of the level of specificity and size of the data set on classification [20]. Our contribution is thus a better understanding of the impact of specificity and the size of the text corpus in word embeddings. Although the results presented here are focused on the legal and Portuguese context, the results and methods can be reused in other applications.

This paper is organized as follows: In Sect. 2, we present some concepts on Brazilian Judiciary organization, word embeddings and text classification. In Sect. 3, we expose some works about word embeddings applied in different

Portuguese domain, and in other languages legal domain. In Sect. 4, we describe the methodology and strategies used in the experiments. In Sect. 5, we show and discuss the results. Finally, there are concluding remarks and new perspectives of work in Sect. 6.

2 Background

To contextualize our dataset and the models used in this work, in the following sections we present relevant concepts on the hierarchy of the Brazilian Judiciary, as well as the tasks of text representation and text classification.

2.1 Courts of Brazilian Judiciary

In order to may correct errors of the judges and also guarantee the non-conformity of the losing party about unfavourable judgments, modern legal systems enshrine the principle of double a degree of jurisdiction. That means that the losing party has the possibility of obtaining a new judgment. For this, all Brazilian Judiciary have higher and lower courts. Above all of them are the Federal Supreme Court (STF), the highest level of the Brazilian Judiciary [13].

According to Brazilian Federal Constitution [1], the Judiciary is composed of: a) Federal Supreme Court (STF); b) Superior Court of Justice (STJ); c) Federal Regional Courts (TRFs) and Federal judges; d) Superior Labor Court (TST), Labor Regional Courts (TRTs) and Labor judges; e) Superior Electoral Court (TSE), Electoral Regional Courts (TREs) and Electoral judges; f) Superior Military Court (STM) and Military judges; g) State Courts (TJs) and State judges. Also, Federal and State Courts can, within their jurisdiction, create the Special Courts (JECs and JEFs), which are responsible for judging local less complex cases. We illustrate this organization in Fig. 1.

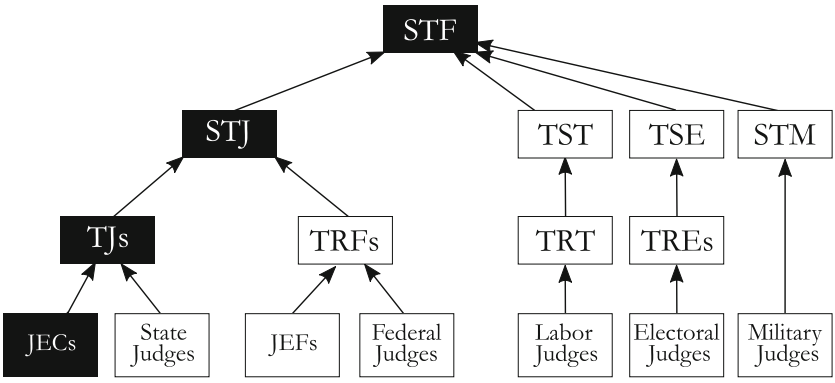

Fig. 1. Brazilian Judiciary organization chart

In our evaluation step, we classify JECs judgments about air transport failures, which belongs to Consumer Law. This legal subject is under the jurisdiction of the State Courts. Therefore, as highlighted in Fig. 1, JECs are submitted to TJs, which in turn are submitted to STJ and STF. So, we used judgments from TJs, STJ and STF to training the word embeddings model.

2.2 Text Classification with CNN

In the text classification task, we use data to construct a model that learns to relate its features to one of the class labels preset. For a given test instance for which the class is unknown, the training model is used to predict a class label for this instance [4].

To evaluate if the prediction is correct (true positives and true negatives) or wrong (false positives and false negatives), we use some metrics, such as accuracy, precision, recall, and F1 score [3]:

- **Accuracy** is the fraction of test instances in which the predicted value matches the ground-truth value.
- **Precision** is the percentage of instances predicted to belong to the positive class that was correct.
- **Recall** is the percentage of ground-truth positives that have been recommended as positives.
- **F1** score is the harmonic mean between the precision and the recall.

When dealing with multiclass datasets we can measure the F1 score using the macro F1 score. It first calculates the metric for each class and then takes the average of them. In this way, the metric considers the performance equally in each class, surpassing any class imbalances. In contrast, there are other average methods such as micro and weighted F1 score however, these variations do not take class imbalance into account [35].

Recently, deep learning models have been proven to be effective in text classification. A popular deep learning model that attracts more attention for text data is the Convolutional Neural Networks (CNNs). CNNs use convolutional masks to sequentially convolve over the data. For texts, a simple mechanism is to recursively convolve the nearby lower-level vectors in the sequence to compose higher-level vectors. Similar to images, such convolution can naturally represent different levels of semantics shown by the text data [23]. This can be better achieved when using text representations where words with similar meanings have similar vectors, as it occurs with Word Embeddings representation [6].

2.3 Text Representation with Word Embeddings

Word embeddings, also known as distributed word representations, can capture both the semantic and syntactic information of words while representing them as n-dimensional dense vectors [20].

These representations are generated from large unlabeled corpora through a training process that varies among existing algorithms. In Word2Vec Skipgram for instance, the representations are generated and modified as it tries to predict surrounding words in a phrase, based on the current word [22]. On the other hand, GloVe creates a co-occurrence matrix containing the frequencies of words in different contexts. Then it applies a dimensionality reduction technique to produce final representations [24].

Word embeddings have been used in many NLP tasks beyond text classification. These tasks include clustering [3], text summarization [5], and many others.

3 Related Work

After a systematic review, we select six works related to ours: a) two about word embeddings applied in multi themes in Portuguese (Brazilian and European), b) two about word embeddings applied in legal theme in several languages, and c) two about text classification in STF.

Hartmann et al. [15] evaluated different word embedding models trained on a sizeable Portuguese corpus (1,395,926,282 tokens in total). They trained 31 word embedding models using FastText, GloVe, Wang2Vec and Word2Vec, and evaluated them intrinsically on syntactic and semantic analogies and extrinsically on POS tagging and sentence semantic similarity tasks. The results obtained from intrinsic and extrinsic evaluations were not aligned with each other, contrary to what they expected. GloVe produced the best results for syntactic and semantic analogies, and the worst, together with FastText, for both POS tagging and sentence similarity.

Rodrigues et al. [25] evaluated different word representation models on semantic similarity tasks, trained on a Portuguese dataset provided by a workshop (10,000 sentences). They used word embeddings (Word2Vec and FastText) and deep neural language models (ELMo and BERT). The results indicated that ELMo language model was able to achieve better accuracy than any other pretrained model which has been made publicly available for the Portuguese language. They also demonstrate that FastText skip-gram embeddings can have a significantly better performance on semantic similarity tasks.

Chalkidis and Kampas [10] trained a word embeddings model on a large legal corpus from various public legal sources in English (UK legislation, European legislation, Canadian legislation, Australian legislation, English-translated legislation from EU countries, English-translated legislation from Japanese, US Supreme Court decisions and US Code), which sums up to a total of approximately 492,000,000 tokens. They trained based on the Word2Vec and Skip-gram model, instead of the most recent FastText. They justified that Word2Vec is reported to provide better semantic representation than FastText, which tends to be highly biased towards syntactic information.

Smywiński-Pohl et al. [29] trained word embeddings models (Word2Vec and GloVe) to find out which of them is best suited for establishing the

correspondence between Polish legal and extra-legal terminology. The corpora are composed of text data collected from two databases: a) National Corpus of Polish, which includes texts of different genres, such as novels, transcripts of parliamentary speeches and newspaper articles, which sums up to a total of 2,591,817,208 tokens; b) judgments from Polish Supreme Court, Polish Constitutional Tribunal, Polish common courts, Polish National Chamber of Appeal and Polish administrative courts, which sums up to a total of 4,076,628,858 tokens. The results showed the superiority of Word2Vec CBOW negative sampling variant in their problem.

Finally, Correia da Silva et al. [28] and Braz et al. [8], representing a national project developed at the STF, classified different types of judgments using deep learning models (Bi-LSTM and CNN) and obtained satisfactory results. However, the published papers suggest that they used a model of word embeddings already trained for the task of text representation.

In our survey, we did not find publications concerned with the training of word embeddings in Portuguese legal texts. Therefore, with this paper, we plan to contribute in this direction.

4 Experiments

To answer our research question, we build different embeddings (varying the specificity and size of the text corpus used to train them) and evaluate their performance in a classification problem. The classification problem concerns specific texts in the area of air transport services. We expect that more specific embeddings require smaller corpus size than general embeddings.

In this section, we explain the pipeline of this work, starting from corpora construction for word embeddings training and also the dataset used for text classification. Then, we describe the embedding training steps and the classification model used to evaluate our embeddings.

4.1 Corpus Construction

The following sections describe the steps[1] we followed for the construction of the text corpora used for (i) training the different word embeddings we want to evaluate as well as (ii) the corpus considered for the text classification.

Concerning embeddings training, the first step is to obtain the collection of legal documents from the court web portals, followed by raw text extraction from these documents. To enable us to evaluate the specificity influence of these legal corpora, we divided it into two contexts: related to general legal texts and related to air transport services text. We also collected texts from other general topics (not related to legal domains) that are already compiled and freely available. Having the corpora for legal and miscellaneous contexts, we applied

[1] Code and Word Embeddings available at https://github.com/thiagordp/embeddings_in_law_paper.

some processing steps to remove noise from texts. To evaluate the influence of corpus size in embeddings training, we divided these three corpora into smaller pieces based on word count.

Concerning the classification task, the construction of the corpora is based on JEC processes related to air transport services. We are thus interested in the quality of the classification in this specific domain and, of course, the impact of the specificity of the embeddings in this specific distribution problem.

Legal Context Corpus for Embeddings Training. To train the embeddings it is required large text corpora to be able to get good embeddings. However, in the Brazilian Portuguese language, we could not find any dataset available on the Internet containing enough legal text corpora for our purposes. Thus, we had to build our legal corpora.

Our main sources of legal text are Brazilian courts platforms. We collected judgments from the webpages of STF [30], STJ [31], and TJ-SC (State Court of Santa Catarina) [34]. We also collected judgments from the JusBrasil portal containing processes related only to failures on air transport service from all TJs (State Courts) of Brazil [16]. Table 1 shows the number of processes acquired and word count for each Tribunal:

Table 1. Acquired process from courts for embeddings training

Source	Collegial judgments	Individual judgments	Subtotal	Word count
STF	64,779	118,910	183,689	294,937,185
STJ	101,141	0	101,141	312,687,450
TJ-SC	989,964	662,535	1,652,499	3,060,212,814
TJs (JusBrasil)	34,239	0	34.239	78,138,337
		Total	1,971,568	3,745,975,786

After downloading all processes, most of them in PDF and Rich Text Format (RTF) formats, we extracted raw texts from these files. We did not apply Optical Character Recognition (OCR) in scanned PDF documents, due to time limits to finish the experiments, so only digital PDFs were accounted with RTF files in Table 1.

With the extracted texts, we applied some pre-processing steps, as discussed further in this section. Then we built the legal text corpora containing all the processes related to all law subjects, which we call *general* legal text corpora in this work. Using this base, we created another text corpora whose processes are related only to air transport and consumer law, and we call it *air transport* legal text corpora.

Global Context Corpus Acquisition. To be able to compare how good embeddings trained with legal texts perform against those created with all kinds

of texts, we also created other corpora from a variety of sources. Thus, we searched for free available textual datasets. In this work, we call these texts as *global* context texts. Table 2 shows all the global text datasets used. Then we apply some preprocessing steps, as will be described further in this section.

Table 2. Global context corpora

Dataset	Documents	Word count	Source
Wikipedia Dump in Portuguese	1,014,713	303,622,360	[2]
Brazilian Literature Books	169	37,848,783	[33]
Old Newspapers	617,627	26,441,581	[32]
Folha de São Paulo News	165,641	74,594,367	[21]
HC News Corpus	494,128	27,170,063	[12]
Blogspot Posts	2,181,073	696,657,915	[26]
Wikihow Instructions	786,283	22,471,312	[11]
Total	5,259,634	1,188,806,381	

Legal Context Corpus for Text Classification. To evaluate each of the trained embeddings, we used a set of judgments from the JEC located at the Federal University of Santa Catarina (JEC/UFSC), which is related only to failures on air transport services (Consumer Law). In these processes, the consumer claims for compensation for material or moral damages against an airline company due to failures in its services. We extracted nearly one thousand judgments, divided into four class labels, corresponding to 26%, 10%, 62%, and 2% of this dataset, respectively:

1. Well-founded: The consumer wins the lawsuit.
2. Not founded: The consumer loses the lawsuit.
3. Partly founded: The consumer wins part of the lawsuit (for example, when he/she plead a greater compensation than the assigned value by the judge).
4. Dismissed without prejudice: The consumer makes a procedural error (for example, when he/she indicate as a defendant the wrong airline company). So the consumer can file a new lawsuit.

Before the text classification task, we applied some preprocessing steps as discussed further in this section and then created three subsets of processes, the training, validation and test sets, corresponding to 70%, 15%, and 15% of the dataset. In these sets, all the classes are distributed proportionally. We used the training and validation sets during the training of the classification model. Then, we evaluated this model with the test set.

Corpus Processing. After text extraction from the documents, we applied some pre-processing steps, which are required before training the embeddings or text classification. The first of them was the conversion to lower case. Then punctuation marks were removed, as well as special characters and some symbol characters. We removed stopwords neither apply stemmization or lemmatization, following the literature [22,24].

In relation to our three corpora used in embeddings training, which comprising 3.7 billion, 100 million and 1.19 bilion words for *general*, *air transport* and *global* corpora, respectively, we created others based on them according to the following smaller corpora sizes, considering the word count: 1,000; 10,000; 50,000; 100,000; 200,000; 500,000; 1,000,000; 5,000,000; 10,000,000; 25,000,000; 100,000,000; 500,000,000; 750,000,000 and 1,000,000,000.

We choose these corpora sizes to be able to compare the variation on evaluation metrics while increasing corpora size. For the air transport context, we could not embrace all these sizes due to limited corpora available. The largest sub-base had 100 million words for this context.

Finally, each of these smaller corpora was used to train one different word embeddings representation.

4.2 Embeddings Training

In this work, we chose GloVe representation due to its good results in many NLP tasks, including text classification, and also for its training time which is significantly less than other techniques like Word2Vec and FastText [24]. In terms of GloVe parameters, we kept most of the default values, except for windows size, training iterations, and vector size, which were set to 5, 100, and 100, respectively. With these values, we achieved better results in text classification.

Considering the corpus sizes described in Sect. 4.1 and the parameters above described, we trained 15 representations for *general* and *global* contexts bases. For *air transport* context base, we trained 11 embeddings.

4.3 Embeddings Evaluation in Legal Text Classification

To evaluate the GloVe embeddings representations, we applied each of them to the task of text classification on judgments from JEC/UFSC. Also, we used Convolutional Neural Networks as a classification model based on the literature [17]. Figure 2 illustrates this model.

This CNN takes into account the order of the words by stacking the corresponding embeddings for each word as they occur in the text. Them it applies multiple convolutional masks with different dimensions that correspond to the red and yellow contours in Fig. 2. Mask widths are equal to word embedding size while the heights can vary. In this context, mask height can be related to the idea of N-Grams, since they embrace multiple embeddings at the same time. In the original model, these heights were set to 3, 4, and 5. We added one more mask of height 2, which increased classification metrics. Also, we set to 10 the number

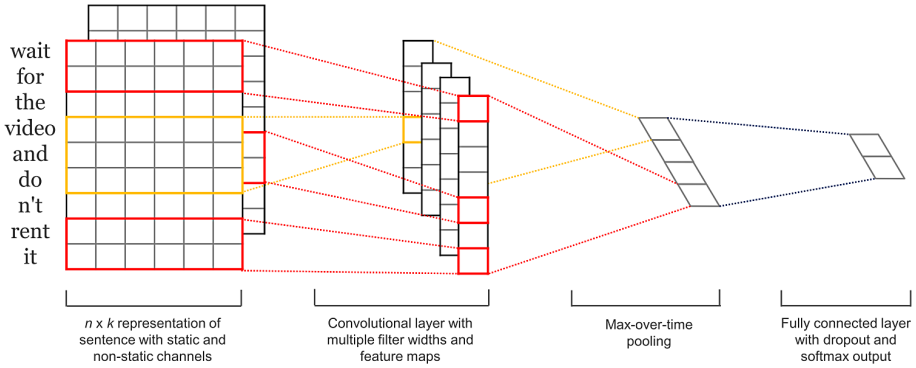

Fig. 2. CNN model for text classification [17]

of masks for each of these sizes, without affecting our results, but decreasing the required training time.

In this work, we applied each of the embeddings trained in conjunction with the CNN described in the classification of JEC/UFSC judgments, where Out of Vocabulary (OOV) words are replaced by an vector of random values. Thus, we trained and tested 41 models. Furthermore, due to the stochastic nature of neural networks training methods [14], each of these models was trained and tested 200 times and the resulting evaluation metrics were averaged.

Finally, we compare the performance in classification using Accuracy and Macro F1-Score.

5 Results Analysis and Discussion

In the following sections, we present and discuss our results for text classification using trained embeddings for *global*, *general*, and *air transport* contexts with multiple corpus training sizes.

5.1 Experimental Results

Following the steps presented in Sect. 4, we trained all 41 word embeddings representations for GloVe.

To illustrate how these embeddings behave, in Fig. 3, we used Principal Component Analysis (PCA) to create a projection in two dimensions of a set of words from *general* context embedding trained with 1 billion words.

Using each embedding, we trained and tested CNNs for text classification in JEC/UFSC judgments. These two steps were repeated 200 times, and the evaluation metrics were averaged for each group of repetitions.

In Fig. 4 and 5, we present the results, for accuracy and F1-Score, respectively, from test data applied to each CNN model. These results are related to embeddings trained with *general*, *air transport*, *global* texts. The x-axis denotes the

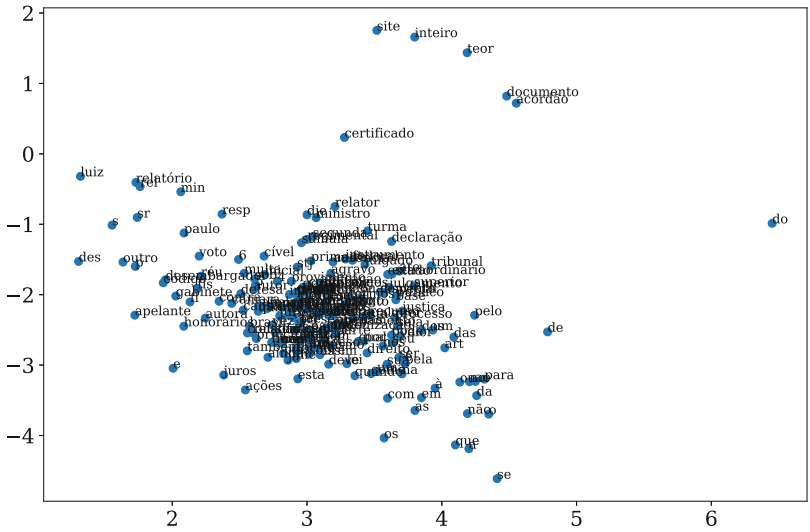

Fig. 3. Word embeddings projection

corpus sizes used to train the embeddings, while the y-axis represents accuracy or F1-Score. Each data point represents the average of the evaluation metric, after 200 train and test repetitions using each specific embedding.

5.2 Discussion from Context Perspective

In this section, we will consider the first part of our research question: Does the specificity of the corpora in embeddings training contribute to the quality of the classification?

In terms of accuracy, when we compare *global* against others (Fig. 4), we have that higher text specificity leads to better results, for most of the corpus sizes used for embeddings training. Furthermore, when comparing *general* and *air transport* curves, there is a significant difference in accuracy only for the lowest and highest x-values. However, in terms of F1-Score, as shown in Fig. 5, our observations change, once *general* and *air transport* curves have a similar shape. Also, for the highest corpus sizes, *general* and *global* curves converge to similar values of F1-Score. We believe that these differences in accuracy and F1-Score emerge from the fact that our dataset to text classification is imbalanced, once the former does not take this fact into account, while the latter does. However, this result still requires further investigation.

In general, we can note that for smaller corpora size for embeddings training, text specificity has a more impact than for large sizes.

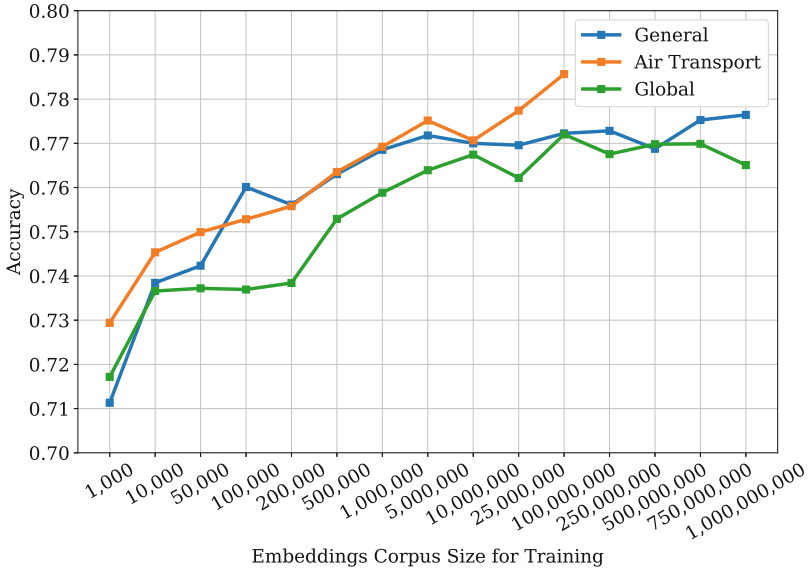

Fig. 4. Accuracy for test set from CNN model

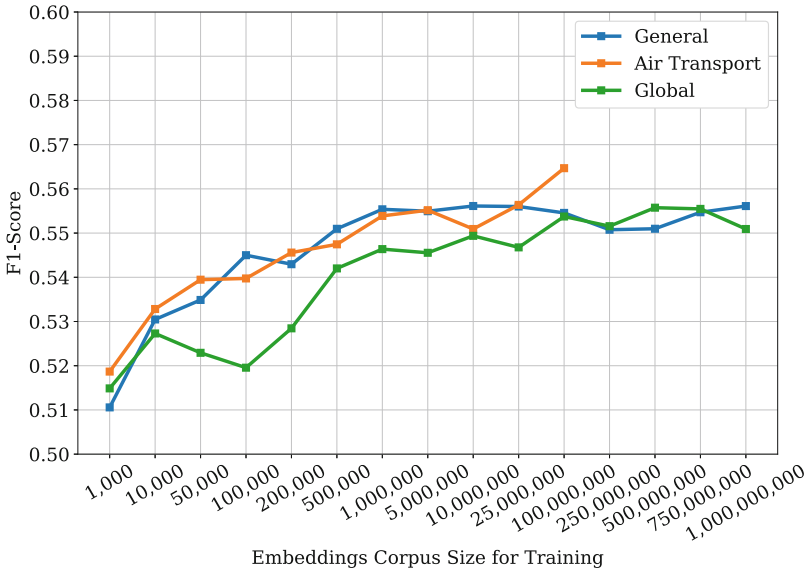

Fig. 5. Macro F1-Score for test set from CNN model

5.3 Discussion from Corpus Size Perspective

In this section, we will consider the second part of our research question: How does the corpus size contribute to the embedding quality?

When we observe both accuracy and F1-Score measures from Fig. 4 and 5, it is clear the tendency for improvement while increasing corpus size. However, the metrics converge with the largest corpus sizes. There are two exceptions. The first one occurs with smaller values of corpus sizes for *global* curve, as it decreases in F1-Score measures. The second corresponds to the last data point in *air transport* curves. The former can happen we the classifier performs poorly for some classes while gets better in others. The latter may indicate that those curves could improve if we had more significant corpus sizes related to that context.

In general, we can note that the greater the corpus size in embeddings training, the better are the results. However, this impact decreases as the corpus size increases until a point where more words in the corpus have little impact on the results.

6 Conclusion and Future Work

The research allowed us to learn more about the behaviour of word embedding models in different variations of the text in the legal domain, such as the specificity (context) and the corpus size.

In the context of legal documents in Portuguese, we concluded that there is more assertiveness when the trained text resembles the text that we have to classify. This behaviour does not occur with the size of the training corpus, because when you reach a certain amount of words, the results suggest stability.

Despite the moderate gain in accuracy with the specific texts as test set (air transport curve), we consider this result relevant because it shows that the use of billions of tokens, as in previous works, does not bring great contributions. Therefore, the specificity of the text set impacts more positively on the classification task than the size of the text set.

Finally, the results presented in this work cover only CNNs. Thus, we intend to check how our conclusions fit when using other classification and representation techniques, although we believe the results would not change significantly. In our future work, we intend to use as classification models SVM, LSTM, Attention Mechanisms etc. Also, we plan to create and experiment new legal datasets to the classification task. We also aim to use other word embeddings models, such as Word2Vec and FastText, as well as other text representation approaches, like BERT, ELMo etc. Finally, we would experiment word embeddings in other tasks, such word analogies and word similarity.

References

1. Brazilian Federal Constitution (1988). http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

2. Ptwiki dump progress on 20191120 (2019). http://wikipedia.c3sl.ufpr.br/ptwiki/20191120/
3. Aggarwal, C.C.: Machine Learning for Text, 1st edn. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73531-3
4. Aggarwal, C.C., Zhai, C. (eds.): Mining Text Data, 27th edn. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4
5. Alami, N., Meknassi, M., En-nahnahi, N.: Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. Expert Syst. Appl. **123**, 195–211 (2019)
6. Aubaid, A.M., Mishra, A.: Text classification using word embedding in rule-based methodologies: a systematic mapping. TEM J. **7**(4), 902–914 (2018)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2016)
8. Braz, F.A., et al.: Document classification using a Bi-LSTM to unclog Brazil's supreme court. In: NeurIPS Workshop on Machine Learning for the Developing World (ML4D), 8 December 2018
9. Cardoso, E.F., Silva, R.M., Almeida, T.A.: Towards automatic filtering of fake reviews. Neurocomputing **309**, 106–116 (2018)
10. Chalkidis, I., Kampas, D.: Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif. Intell. Law **27**(2), 171–198 (2019). https://doi.org/10.1007/s10506-018-9238-9
11. Chocron, P., Pareti, P.: Vocabulary alignment for collaborative agents: a study with real-world multilingual how-to instructions. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, pp. 159–165, July 2018
12. Christensen, H.: HC Corpora (2016). https://web.archive.org/web/20161021044006/http://corpora.heliohost.org/
13. Cintra, A.C.d.A., Grinover, A.P., Dinamarco, C.R.: Teoria geral do processo. Malheiros (2011)
14. Cohen, P.R.: Empirical Methods for Artificial Intelligence. MIT Press, Cambridge (1995)
15. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks (Section 3), August 2017
16. JusBrasil: JusBrasil. Conectando pessoas à justiça (2020). https://www.jusbrasil.com.br/home
17. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 2017-January, pp. 1746–1751. Association for Computational Linguistics, Stroudsburg, September 2014
18. Kowsari, K., Meimandi, J., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: a survey. Information **10**(4), 150 (2019)
19. Kumar, G.R., Mangathayaru, N., Narasimha, G.: Intrusion detection using text processing techniques. In: Proceedings of the The International Conference on Engineering & MIS 2015 - ICEMIS 2015. ACM Press (2015)
20. Lai, S., Liu, K., He, S., Zhao, J.: How to generate a good word embedding. IEEE Intell. Syst. **31**(6), 5–14 (2016)
21. Marlessonn: News of the Brazilian newspaper (2019). https://www.kaggle.com/marlesson/news-of-the-site-folhauol

22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, pp. 1–12, January 2013
23. Peng, H., et al.: Large-scale hierarchical text classification with recursively regularized deep graph-CNN. In: Proceedings of the 2018 World Wide Web Conference, WWW 2018, pp. 1063–1072. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva (2018)
24. 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), vol. 19, pp. 1532–1543. Association for Computational Linguistics, Stroudsburg (2014)
25. Rodrigues, R.C., Rodrigues, J., de Castro, P.V.Q., da Silva, N.F.F., Soares, A.: Portuguese language models and word embeddings: evaluating on semantic similarity tasks. In: Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., Gonçalves, T. (eds.) PROPOR 2020. LNCS (LNAI), vol. 12037, pp. 239–248. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-41505-1_23
26. Santos, H., Woloszyn, V., Vieira, R.: BlogSet-BR: a Brazilian Portuguese blog corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki (2018)
27. Sheikhalishahi, S., Miotto, R., Dudley, J.T., Lavelli, A., Rinaldi, F., Osmani, V.: Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med. Inform. **7**(2), e12239 (2019)
28. da Silva, N.C., et al.: Document type classification for Brazil's supreme court using a convolutional neural network. In: 10th International Conference on Forensic Computer Science and Cyber Law (ICoFCS), Sao Paulo, Brazil, October 2018
29. Smywiński-Pohl, A., Lasocki, K., Wróbel, K., Strzałta, M.: Automatic construction of a polish legal dictionary with mappings to extra-legal terms established via word embeddings. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL 2019. ACM Press (2019)
30. STF: Supremo Tribunal Federal (2020). http://portal.stf.jus.br/
31. STJ: STJ - Jurisprudência do STJ (2020). https://scon.stj.jus.br/SCON/
32. Tan, L.: Old newspapers (2020). https://www.kaggle.com/alvations/old-newspapers
33. Tatman, R.: Brazilian literature books (2017). https://www.kaggle.com/rtatman/brazilian-portuguese-literature-corpus
34. TJSC: Jurisprudência Catarinense - TJSC (2020). http://busca.tjsc.jus.br/jurisprudencia/
35. Uysal, A.K.: An improved global feature selection scheme for text classification. Expert Syst. Appl. **43**, 82–92 (2016)
36. Wang, S., Zhou, W., Jiang, C.: A survey of word embeddings based on deep learning. Computing **102**(3), 717–740 (2019)