Sentiment Classification over Brazilian Supreme Court decisions using Multi-Channel CNN

Marcus O. Silva[1] - marcus.oli.silva@gmail.com, Gustavo C. Bicalho[1] - gustavocbicalho@gmail.com, Thiago de Paula Faleiros[1] - thiagodepaulo@unb.br, and Henrique Araujo Costa[2] - henriquearcos@unb.br

[1] Department of Computer Science - University of Brasilia - Brazil
[2] Law School - University of Brasilia - Brazil

Abstract. Sentiment analysis seeks to identify the viewpoint(s) underlying a text document; In this paper, We present the use of a multichannel convolutional neural network which, in effect, creates a model that reads text with different n-gram sizes, to predict with good accuracy sentiments behind the decisions issued by the Brazilian Supreme Court, even with a very imbalanced dataset we show that a simple multichannel CNN with little to zero hyperparameter tuning and word vectors, tuned on network training, achieves excellent results on the Brazilian Supreme Court data. We report results of 97% accuracy and 84% average F1-score in predicting multiclass sentiment dimensions. We also compared the results with classical classification machine learning models like Naive Bayes and SVM.

Keywords: Convolutional Neural Network · Sentiment analysis · Natural Language Processing · Document Classification · Jurisprudence.

1 Introduction

Sentiment analysis and text classification methods have been successfully used on many NLP applications and tasks, moreover, the use of AI and Deep Learning for legal text analytics is more and more becoming a reality [3]. In this paper we propose the use of convolutional neural networks to extract sentiment classifications on decisions issued by the Brazilian Supreme Court (STF), this kind of technic can be applied or extended to the decisions of another forign courts. To the best of our knowledge, this approach is relatively new and under-explored in the legal domain.

In this sense, The Brazilian Supreme Court(SFT) system is one of the biggest judiciary systems in the world, and receives an extremely high number of cases, because of that there are a lot of final decisions issued by the court every day [1]. In the end of the processes, the court needs to emit a certificate of judgment, which is a textual document in natural language, containing the ruling of the processes, the possibility to automate the classification of sentiment entities of those texts has a lot of applications and benefits. This approach can leverage and facilitate the legal professional work, and improves data analytics on a unstructured data.

A legal tech research group (DireitoTech) developed a sentiment analysis methodology and built an annotation tool that was used by a team of law professionals who manually classified the sentiment of decisions from those certificates of judgment documents. In this work, we considered four labels from that sentiment classification methodology. The sentiment itself: positive(favorable to the defendant) or negative. And the decision form: unanimous or majority. We grabbed a subset of decisions of the processes of the type "judicial review"(in Portuguese, Recursos extraordinários) and extracted a dataset considering the four classifications described above.

Using this dataset, this paper reports the results of a preliminary evaluation of 3605 documents from the STF manually labeled by the legal professionals. We propose a multichannel convolutional neural network architecture to classify the decisions on these 4 sentiments dimensions and show that it obtains 84% F1-score macro average on a very unbalanced dataset. Moreover, it is shown that the CNN model outperform, with a large margin, others classical machine learning algorithms.

The remainder of this paper is organized as follows. In Section 2 we show the motivation of the work for the law practitioners. In Section 3 we discuss about the related works. In Section 4 we explain the detail about the documents and the dataset. In Section 5, we detail the architecture and multi-channel strategy of our model. Section 6 summarizes the comparative results of our proposed model against the classical machine learning algorithms. Section 7 offers a conclusion and shows future works.

2 Motivation

Legal practitioners uses several forms of legal information, being the main two: the law and jurisprudence. The law is an abstract norm, that means, it has not been appliact to a concrete case. While the jurisprudence is a concrete norm, made to solve a case submited to the Judicial Power.

Eventhough it is relatively easy to know the laws, because they are published in official repositories, it is much more complex to know the jurisprudence. The most used legislative repository used in Brazil, is the Planalto(avaible at official government website[1]) and it illustrates well how the many forms of legislation in Brazil are organized and consumed. In other hand, there are many courts and each one if responsible to publish its own jurisdiction[5].

In general, courts treat these data as documents in natural language, with a relatively limited additional layer of metadata. This way, there few filters to access this information, for exemple: the date of the judgment, the name of the judge, the agency that owns this judge, the name and the role of each part in the proccess [5]. We couldn't find, however, any public organized repository regarding the sentiment dimension of the judged, if it was favorable or unfavorable in the end.

[1] http://www4.planalto.gov.br/legislacao/

It is possible to imagine that a lawyer in a bank needs a research in the jurisprudence of determined court to evaluate if a new case has any chance of obtaning success. The way it is organized today, he can easly find concrete cases that concerns about a certain topic. However, he will have a hard time finding, inside this topic, which were the successful cases for the bank and which the same bank had a defeat.

The utility in developing a solution that comprehend which are the favorable and unfavorable cases is in making feasible an aggregate query also for this sentiment dimension. After all, the professional consultation almost always has an interested side, in a way that knowing the outcome of the case is an information for the practical life of law professionals.

In that context, The DireitoTech[2] research team and a group of law professionals, developed an annotation platform that was used to classify the sentiment dimesions of the STF certificates of judgment documents, allowing collaborative workflow and simultaneous access of the researchers to the data collection. A simplified example of how the mode (majority or unanimous) and the sentiment itself (negative or positive) was evaluated/annotated can be seen in the Figure 1.

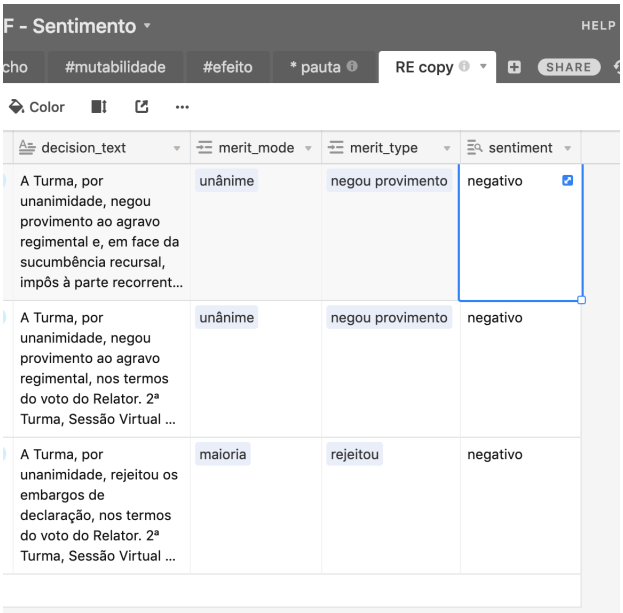

Fig. 1. Plataform used to label sentiment dimensions. Adapted from [5].

[2] https://sites.direitotec.com.br/home

3 Related Works

The paper [4] was the forerunner work of empirical research to apply the techniques of sentiment analysis to documents in legal domain. In particular, this investigation tracked the opinions coming from social media and the Web, that is, from "blogosphere" in legal field. The experiments was based in an application of a existin sentiment analysis tool, LingPipe. This tool is based on document-level classification algorithm. The polarity of each sentences is determined in isolation by standard classification algorithm, as SVM and Naive Bayes. To determine the polarity of a document, it is used a graph representation of sentences polarities, and a graph minimum cut based objective function is formulated to determine the "graph happiness". Despite surpass a baseline of random assignment, the results , with an accuracy and F-score averaging around 60%, remain below pratical requirements.

In the paper [13] it is investigated the application of text classification methods to the legal domain. In it paper, it is used the diachronic collection of court ruling from the French Supreme Court. It is tacked 3 tasks: 1) Prediction the law area of cases and rulings; 2) Predictiong the court ruling; 3) and to estimate the time span when a case description and ruling were issued. The proposed approach was based on classifier ensembles of SVM classifiers. The results of experimental evaluation were compared with results reported in [13] who approach using tradictional SVM classifier trained on bag of words and bag of bigrams. It is showed that a classifier based on SVM ensembles can obtain high scores in prediction the law area and the ruling of a case, given the textual features. The scores are closed to 96% in the task of court ruling prediction, and 96.5% in the task of law area prediction. This demonstrates the feasibility of applying these classification techniques to legal data.

In [11], it was developed a framework to assist professionals in judgement prediction. The method contains two stages: relevant article retrieval through multilabel classification and judgement category forecast based on sentiment analysis. In the stage 1, a standard text classifier model is trained based on space vector representation and SVM classifier. Because each precedent is annotated with multiple labels (i.e. related articles) according to the article classification model, a SVM model generates predictions, leading to an estimated ranking of all articles for an input judgement. In the stage 2, the text is preprocessed and four distinguishing features are selected: sentiment score, punishment period (average period and total period), and top k cited articles (extracted from stage 1). These features are used to create the judgement vector representation of the document. The judgement vector, with regard to judgement label determination, are employed to train a classification model, and to predict a possible category for the judgement. The paper is domain specific and the data set consists of Chinese words, it was employed a corpus-based approach using the Chinese sentiment dictionary, NTUSD [10]. Experimental results from a judgement data set reveal that approach is a satisfactory method for judgement classification, but the experimental analysis was not comprehensive enough to evaluate the performance of several classification algorithms.

4 The dataset

On the STF there types of cases that are dealt by the court, like: Appelate Decision, Extraordinary Appeal, Administrative Orders and others [1]. In this paper, we used the data from the subtype 'Extraordinary Appeal' (in portuguse, 'Recurso Extraordinário')

A total of 3,605 text documents were manually labeled by a team of specialist lawyers. This dataset was split into two parts: 80% of the samples for training, and 20% for test/validatin.

Our work focuses on classifying four main types of sentiments on the Extraordinary Appeal certificates of judgment documents issued by the STF. These are listed in the Table 1, keeping their original label in Portuguese.

Table 1. Dataset sentiments classes.

Label(in portuguese)	Description	Samples	%
negativo_maioria	Negative majority	131	3.7%
negativo_unânime	Negative unanimous	3226	89.5%
positivo_maioria	Positive majority	62	1.72%
positivo_unânime	Positive unanimous	186	5.5%

As we can see, the classes labels are very unbalanced, the vast majority of the documents lay on 'Negative unanimous', and there are little samples of other classes, even with this quality, the proposed model achieves good results, as we can see in the next sections.

To keep the proportion of the classes we stratified the train/test split over the dataset, the final proportion became as follows on the Table 3.

Table 2. Dataset train test samples

Class	Train	Test	%
Negative majority	104	27	3.7%
Negative unanimous	2580	646	89.5%
Positive majority	49	13	1.72%
Positive unanimous	148	38	5.5%

4.1 Preprocessing

To reduce the complexity of the dataset and improve the model's accuracy preprocessing were applied to the documents before extracting features using the following procedures:

– Characters in documents were converted to lowercase.

- Removal of special characters
- Removal of alphanumeric terms with numbers and letters in the same "words".
- Removal of portuguese stop words
- Removal of short tokens
- All documents were tokenized by using NLTK tokenizer.

Text example before preprocessing:

A Turma, por unanimidade, negou provimento ao agravo regimental e, em face da sucumbência recursal, impôs à parte recorrente o pagamento de honorários advocatícios adicionais equivalentes a 20% (vinte por cento) do valor a esse título já fixado noprocesso (CPC/2015, art. 85, § 11), nos termos do voto do Relator. 2ª Turma, Sessão Virtual de 25.11 a 1º.12.2016.

Text example after preprocessing:

turma unanimidade negou provimento agravo regimental face sucumbência recursal impôs parte recorrente pagamento honorários advocatícios adicionais equivalentes vinte cento valor título fixado noprocesso art termos voto relator turma sessão virtual

5 Proposed Method

The literature shows that one of the state-of-the-art approaches for document classification consists in applying a Convolutional Neural Network (CNN) on embedded text [7]. The proposed architecture consists of a multi-channel Convolutional Neural Network with embedding layers. This CNN is a slightly modified version of the architecture used by [8]. This work provided simple and effective architecture for text classification, where convolutional layer can extract local n-gram features.

Figure 2 illustrates the architecture utilized. It is a three channels architecture that can be sepparated in 5 layers: an embedding layer, a convolutional layer, a max-pooling layer, a concatenation layer and finally a sigmoid layer.

5.1 Embedding layer

We used word embedding as an input layer of the convolutional layer. This layer transforms each token/word in a distributed vector of 100 dimensions. The words are randomly initialized and then modified during training.

5.2 Convolutional

The input local n-gram features were extracted by convolutionals. This convolutional layers was added with kernel sizes 4,6,8 respectively for each channel with 32 filters resulting in a output of dimensions: (257, 32), (255, 32), (253, 32).

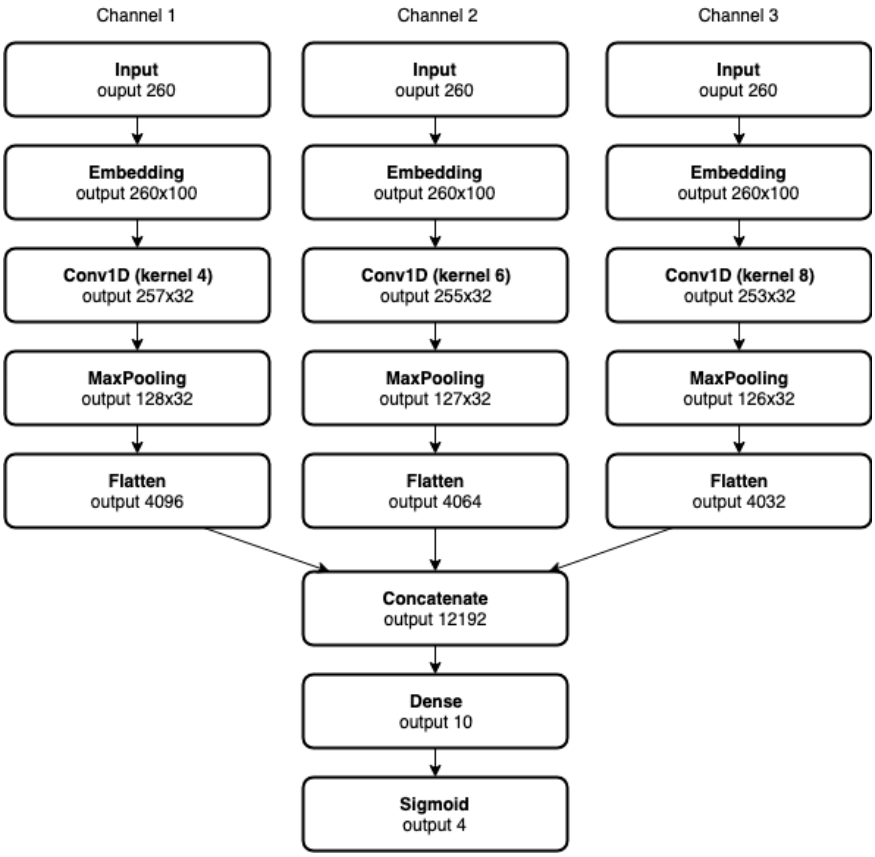

Fig. 2. Multi-channel Convolutional Neural Network.

5.3 Max pooling layers

Then, max pooling over channels was applied to the output of the convolutional layer. The max pooling chooses the part of the data with greater relevance for classification of documents.

5.4 Concatenation layer

In order to get all the channels, we used concatenation layer to merge the three channels, Leading to a one-dimensional array of 12192 dimensions.

5.5 Sigmoid layer

In the sigmoid layer, the outputs of the concatenation layer are converted into classification probabilities. In order to compute the classification probabilities, the sigmoid function was used. The output has four dimensions (positive, negative and neutral classes).

5.6 Loss function

This network was trained using the binary cross-entropy as its loss function, and the Adam optimization method. As explained in Section 3, the dataset is very unbalanced, We added class weight penality in our loss function, using the following proportion:

5.7 Implementation

We used Keras with tensorflow [2] backend, we utilize the CPU computing resource. The optimizer we used was Adam [9]. We also applied dropout(0.5) to each convolutional layer to prevent overfitting. And runned the model for 10 epochs.

Table 3. Class weights

Class	Weight
Negative majority	6.92548077%
Negative unanimous	0.27916667%
Positive majority	14.69897959%
Positive unanimous	4.86655405%

6 Results

To test our model, we experimented other classical classification methods on the same dataset. To select the best parameters for each method, we used the GridSearchCV method from the scikit-learn library for Python [12].

The models we tested and the parameters for each one of them are:

- Logistic Regression - C: 100.0, penalty: l1
- Multinomial Naive Bayes - alpha: 1, fit_prior: True
- Complement Naive Bayes - alpha: 1, fit_prior: True
- SVM - C:10, kernel: linear

As a result from all the tests, we can see on Tables 4, 5 and 6 that the Multi-channel CNN-based method overperforms the classic classification methods in all parameters for all classes. As we have an unbalanced dataset, the performance of the other methodologies are hugely affected, showing really poor results for *negativo_ maioria, positivo_ maioria and positivo_ unânime* classes, while the Multi-channel CNN-based method still has great performance for all of those classes. And, as most of our data comes from *negativo_ unânime*, all models achives over 94% F1-score for that specific class.

When we analyse the F1-score macro average and the model accuracy, the Table7 shows that our Multi-channel CNN-based method outperforms all other

tested methods by a huge margin. While we achieve a F1-score macro average of 84%, the second best method, Logistic Regressio, only achieved 40%. And the difference can also be seen in the accuracy of the models, while we achieve 97% accuracy, the second best, Complement Naive Bayes, only achieved 87%.

Table 4. Precision Table for each individual class.

Precision	negativo_maioria	negativo_unânime	positivo_maioria	positivo_unânime
Multi-CNN	0.91	1.00	0.62	0.82
LogisticRegression	0.27	0.94	0.15	0.26
MultinomialNB	0.00	0.92	0.17	0.20
ComplementNB	0.00	0.91	0.20	0.27
SVM	0.27	0.94	0.00	0.33

Table 5. Recall Table for each individual class.

Recall	negativo_maioria	negativo_unânime	positivo_maioria	positivo_unânime
Multi-CNN	0.74	0.99	0.77	0.95
LogisticRegression	0.30	0.95	0.15	0.18
MultinomialNB	0.00	0.98	0.08	0.11
ComplementNB	0.00	0.99	0.15	0.08
SVM	0.26	0.95	0.00	0.37

Table 6. F1-score Table for each individual class.

F1-score	negativo_maioria	negativo_unânime	positivo_maioria	positivo_unânime
Multi-CNN	0.82	0.99	0.69	0.88
LogisticRegression	0.28	0.94	0.15	0.22
MultinomialNB	0.00	0.95	0.11	0.14
ComplementNB	0.00	0.95	0.17	0.12
SVM	0.26	0.95	0.00	0.35

7 Conclusion

The sentiment dimension of the decisions issued by the Brazilian Supreme Court is an important factor of the jurisprudence study and analysis by the law professionals. In this paper, We proposed a multi-channel CNN with embeddings vector for sentiments classification over those decisions. Our model outperformed other classical classifiers and showed an impressive result in the dataset with imbalanced labels.

Table 7. F1-score Macro Average and Accuracy for each model tested.

Model	F1-score Macro Average	Accuracy
Multi-CNN	0.84	0.97
LogisticRegression	0.40	0.87
MultinomialNB	0.30	0.88
ComplementNB	0.31	0.89
SVM	0.39	0.87

For future work, the application of other neural networks architectures, like CNN-LSTM[14] or CNN-BiLSTM[6], can be applied to improve our model. A grid-search can be used in the future to better tunning hyperparameters on the CNN. In the pipeline, we wish to test and improve our model in a bigger dataset, with more classes, more sentiments dimensions, and more decisions types

References

1. Braz, F.A., da Silva, N.C., de Campos, T.E., Chaves, F.B.S., Ferreira, M.H.S., Inazawa, P.H., Coelho, V.H.D., Sukiennik, B.P., de Almeida, A.P.G.S., Vidal, F.B., Bezerra, D.A., Gusmao, D.B., Ziegler, G.G., Fernandes, R.V.C., Zumblick, R., Peixoto, F.H.: Document classification using a bi-lstm to unclog brazil's supreme court. CoRR **abs/1811.11569** (2018), http://arxiv.org/abs/1811.11569
2. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
3. Conrad, J.G., Branting, L.K.: Introduction to the special issue on legal text analytics. Artificial Intelligence and Law **26**(2), 99–102 (Jun 2018). https://doi.org/10.1007/s10506-018-9227-z, https://doi.org/10.1007/s10506-018-9227-z
4. Conrad, J.G., Schilder, F.: Opinion mining in legal blogs. In: Proceedings of the 11th international conference on Artificial intelligence and law. pp. 231–236. ACM (2007)
5. Costa, H.: Classificando decisões judiciais com inteligência artificial: primeira parte. https://henarcos.com.br/classificando-decisoes-judiciais-com-inteligencia-artificial-primeira-parte/ (2019)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
7. Kidwelly, P. (ed.): Document type classification for Brazil's supreme court using a Convolutional Neural Network. THE TENTH INTERNATIONAL CONFERENCE ON FORENSIC COMPUTER SCIENCE AND CYBER LAW - ICoFCS, The name of the publisher (2018). https://doi.org/10.5769/C2018001, http://dx.doi.org/10.5769/C2018001
8. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://dblp.uni-trier.de/db/journals/corr/corr1412.htmlKingmaB14
10. Ku, L.W., Chen, H.H.: Mining opinions from the web: Beyond relevance retrieval. Journal of the American Society for Information Science and Technology **58**(12), 1838–1850 (2007)

11. Liu, Y.H., Chen, Y.L.: A two-phase sentiment analysis approach for judgement prediction. Journal of Information Science **44**(5), 594–607 (2018)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
13. Sulea, O.M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., van Genabith, J.: Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306 (2017)
14. Zhang, H., Wang, J., Zhang, J., Zhang, X.: Ynu-hpcc at semeval 2017 task 4: Using a multi-channel cnn-lstm model for sentiment classification. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 796–801. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/S17-2134, http://aclweb.org/anthology/S17-2134