Towards Automated Classification of Repetitive Themes in Brazilian Courts with *LegalClass*[⋆]

Daniela L. Freire[1][0000−0002−5363−3608],
Alex M. G. de Almeida[2][0000−0002−6805−3753],
Márcio de S. Dias[3][0000−0003−1116−6965],
Adriano Rivolli[4][0000−0001−6445−3007],
Fabíola S. F. Pereira[5][0000−0003−2914−1803],
Giliard A. de Godoi[4][0000−0002−1715−0852], and
Andre C. P. L. F. de Carvalho[1][0000−0002−4765−6459]

[1] University of Sao Paulo, Sao Paulo, Brazil {danielalfrere,andre}@icmc.usp.br
[2] Ourinhos College of Technology, Brazil
alex.marino@fatecourinhos.edu.br
[3] Federal University of Catalan, Brazil
marciodias@ufcat.edu.br
[4] Federal Technological University of Paraná, Brazil
rivolli@utfpr.edu.br giliardgodoi@alunos.utfpr.edu.br
[5] Federal University of Uberlândia, Brazil
fabiola.pereira@ufu.br

Abstract. The growing influx of lawsuits in judicial systems presents a pressing challenge for timely case resolution. The Sao Paulo Justice Court is particularly noteworthy, boasting the world's largest caseload with an 84% congestion rate and an average processing time of over seven years. To address this issue, this article introduces *LegalClass*, a computational tool designed to expedite case processing. *LegalClass* employs natural language processing and an array of machine learning algorithms—such as Support Vector Machines, Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Convolutional Neural Networks—for the automated classification of lawsuits, with a focus on repetitive legal themes. Developed in collaboration with the University of São Paulo, this tool aims to significantly enhance the efficiency of the judicial system by harnessing the capabilities of artificial intelligence and data science. This study uses *LegalClass* to assess the performance of machine learning algorithms in categorizing lawsuits according to predefined themes set by the Superior Court of Justice. While automation through text classification has shown promise in handling vast volumes of legal texts, ongoing improvements in methods and techniques are essential for increasing both efficiency and accuracy. Continued research in this rapidly evolving field is crucial for meeting the changing needs of legal information processing.

[⋆] Supported by Tribunal de Justiça de São Paulo

Keywords: Lawsuit classification · Repetitive themes · Natural language processing · Machine learning · Artificial intelligence · Jurisdictional provision.

1 Introduction

The global litigation landscape is witnessing an unprecedented surge, with the Sao Paulo Justice Court (Tribunal de Justiça de São Paulo - TJSP) leading in case volume. This court not only has the highest number of lawsuits worldwide but also has an alarming congestion rate of 84% [9]. Moreover, it has the longest average case processing time among state-level courts in Brazil, stretching to seven years and five months. The sheer volume of paperwork burdens court staff considerably, making it imperative to seek automated solutions to accelerate case resolutions.

In the Brazilian legal system, the concept of Repetitive Appeal offers a streamlined approach to handling multiple special appeals involving identical legal disputes. Selected cases are escalated to the Superior Court of Justice (Superior Tribunal de Justiça - STJ) for adjudication, with related issues in abeyance. Once the court resolves the repetitive appeal, its ruling becomes a canonical "Repetitive Theme," applicable to all suspended cases.

Recognizing the need for technological interventions, the University of São Paulo (USP) has collaborated with TJSP to foster research and development in artificial intelligence, specifically in data science for legal applications. This paper introduces *LegalClass*, a computational tool leveraging natural language processing and machine learning algorithms to analyze and categorize court decisions. The algorithms include Support Vector Machines [4], Logistic Regression [6], Naive Bayes [8], K-Nearest Neighbors [2], and Convolutional Neural Networks. *LegalClass* aims to classify lawsuits into STJ's Repetitive Themes, providing individual and batch processing options and statistics for previously classified cases.

Although various tools employing machine learning for classification tasks exist in the literature [3,5,11], there is a conspicuous absence of text classification solutions explicitly designed for the legal domain in the Portuguese language. This gap in research accentuates the novelty and necessity of *LegalClass*.

Despite the promise of automated text classification, the field is still in its infancy and requires more robust, accurate methodologies to help practitioners rapidly sift through critical information. As such, this paper contributes to ongoing research to refine text classification techniques for enhanced decision-making and productivity.

The remainder of this paper is organized as follows: Section 2 details the proposed tool, Section 3 recounts its application to a real-world case study, and Section 4 concludes with a summary and directions for future research.

2 Proposed Tool

This section outlines the functionality of our web-based tool, *LegalClass*, designed to evaluate machine learning algorithms for classifying Portuguese legal cases into STJ repetitive themes.

2.1 Training

The initial version of *LegalClass* can classify legal cases into five predefined repetitive themes: 0929, 1015, 1033, 1039, and 1101. Table 1 summarizes each theme's focus. Our dataset consists of 10,684 legal decision texts distributed across these themes as follows: theme 0929 has 2,065 samples, theme 1015 comprises 1,237 samples, theme 1033 contains 5,356 samples, theme 1039 includes 526 samples, and theme 1101 has 1,500 samples. The data were partitioned into an 85% training set and a 15% validation set.

Table 1. Summarized Repetitive Theme Descriptions

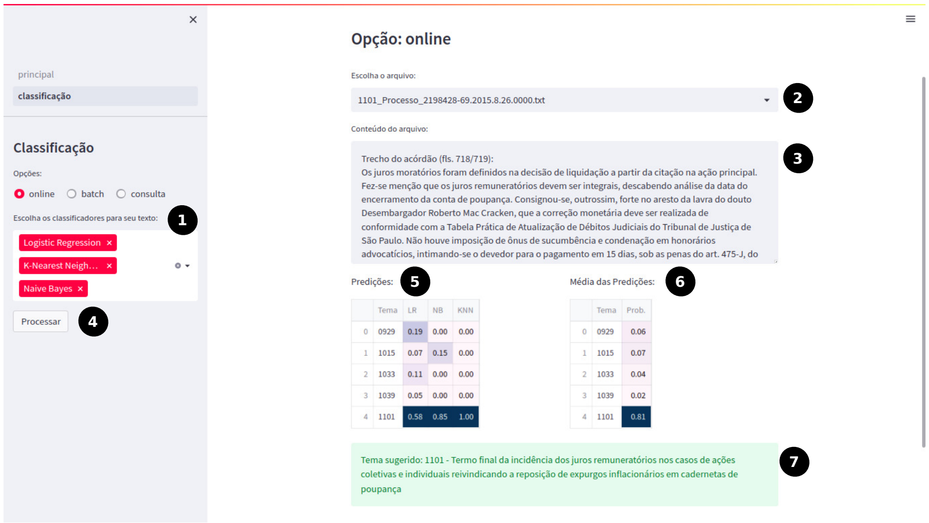
Theme	Concise Descriptions
0929	Pertains to double repetition rules under Consumer Protection Code Article 42.
1015	Involves the financial responsibility of HSBC Bank Brasil S/A due to its business succession with Banco Bamerindus S/A, particularly concerning inflationary adjustments to savings accounts.
1033	Addresses the reset of the statute of limitations for compliance claims related to collective rulings due to the filing of protest actions.
1039	Focuses on the determination of the initial term for indemnity claims against insurers in active or expired Housing Financial System contracts.
1101	Examines the final term for the applicability of interest rates in actions related to inflationary adjustments in savings accounts.

We removed stop words and punctuation marks for text preprocessing and performed specific term transformations relevant to the Brazilian legal context. The Natural Language Toolkit (`nltk`) and `string` libraries facilitated this preprocessing step. The texts were then vectorized using the TF-IDF technique, generating a feature set of 64,687 bi-grams and tri-grams. We employed several machine learning algorithms for multi-class classification, including a Calibrated Classifier with Support Vector Machines (SVM), Logistic Regression (LR), Multinomial Naive Bayes (NB), and K-Nearest Neighbors (KNN).

2.2 User Interface

The *LegalClass* tool offers a user-friendly web interface with three primary functions: online classification, batch classification, and query of past classifications. Built using the open-source *Streamlit* library (version 1.10.0), the tool provides real-time, scalable solutions for classifying legal texts.

According to Fig.1, in the online option, it is possible to select one or more classifiers in (1) and select the document text to be classified in (2). The document text chosen is shown in (3). By the "Process" button in (4), it is possible to execute the classification model code. As a result, two tables are displayed. The former (5) shows the probabilities of each theme of each classifier. The latter (6) shows the average of the probabilities of each theme. The theme is suggested, and its description is displayed in (7).



Opção: online

Escolha o arquivo:

1101_Processo_2198428-69.2015.8.26.0000.txt

Conteúdo do arquivo:

Trecho do acórdão (fls. 718/719):
Os juros moratórios foram definidos na decisão de liquidação a partir da citação na ação principal. Fez-se menção que os juros remuneratórios devem ser integrais, descabendo análise da data do encerramento da conta de poupança. Consignou-se, outrossim, forte no aresto da lavra do douto Desembargador Roberto Mac Cracken, que a correção monetária deve ser realizada de conformidade com a Tabela Prática de Atualização de Débitos Judiciais do Tribunal de Justiça de São Paulo. Não houve imposição de ônus de sucumbência e condenação em honorários advocatícios, intimando-se o devedor para o pagamento em 15 dias, sob as penas do art. 475-J, do

Predições:

	Tema	LR	NB	KNN
0	0929	0.19	0.00	0.00
1	1015	0.07	0.15	0.00
2	1033	0.11	0.00	0.00
3	1039	0.05	0.00	0.00
4	1101	0.58	0.85	1.00

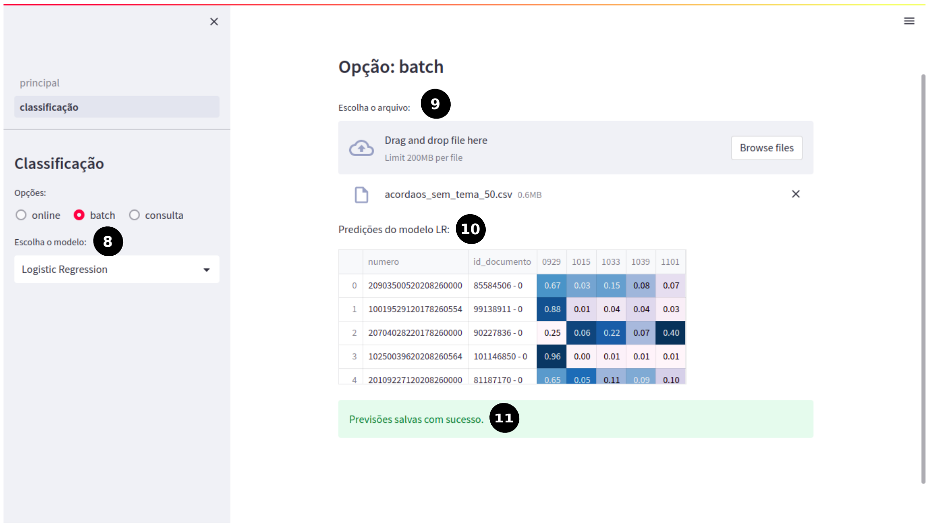
Média das Predições:

	Tema	Prob.
0	0929	0.06
1	1015	0.07
2	1033	0.04
3	1039	0.02
4	1101	0.81

Tema sugerido: 1101 - Termo final da incidência dos juros remuneratórios nos casos de ações coletivas e individuais reivindicando a reposição de expurgos inflacionários em cadernetas de poupança

Fig. 1. Online Classification Interface.

According to Fig.2, in the batch option, it is possible to select one single classifier in (8) and download the file in ".csv" format with the document text list to be classified in (9). The result of classification is another list with document text identifier and their respective probabilities of bellowing to each repetitive theme. This result is shown in (10), where the probabilities are displayed in a heatmap, where the highest probabilities are the darkest. This result is also saved in a folder, and this information is presented in (11). According to Fig.3, in the query option, it is possible to select a list of previously classified document texts in (12). This list, shown in (13), is a heatmap with document text identifier and their respective probabilities. By the "See statistics" button in (14), it is possible to obtain some information about the document text set and their probabilities, such as count, mean, standard deviation (std), minimal, statistical quartiles(25%, 50%, 75%), and maximal, shown in (15). In (16), it is shown the amount (we used the symbol "#") and the percentage (we used the symbol "%") of samples with probabilities greater than or equal to 50%, %, 70%, 80%, and 90%. The



Predições do modelo LR:

	numero	id_documento	0929	1015	1033	1039	1101
0	20903500520208260000	85584506 - 0	0.67	0.03	0.15	0.08	0.07
1	10019529120178260554	99138911 - 0	0.88	0.01	0.04	0.04	0.03
2	20704028220178260000	90227836 - 0	0.25	0.06	0.22	0.07	0.40
3	10250039620208260564	101146850 - 0	0.96	0.00	0.01	0.01	0.01
4	20109227120208260000	81187170 - 0	0.65	0.05	0.11	0.09	0.10

Previsões salvas com sucesso.

Fig. 2. Batch Classification Interface.

latter (6) shows the average of the probabilities of each theme. The theme is suggested, and its description is displayed in (7).

3 Case Study

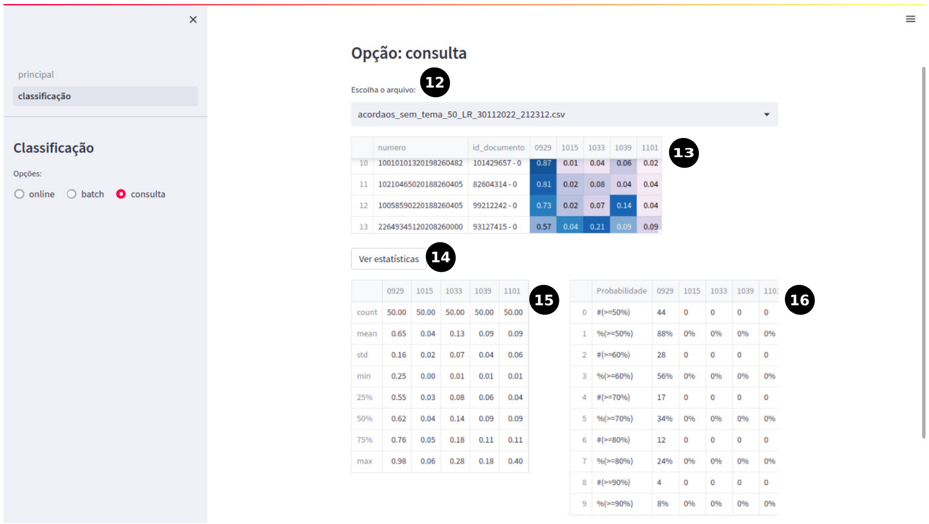
This section outlines an experiment using *LegalClass* to categorize lawsuit decisions. We employ four machine learning classifiers and follow the protocols established in previous software engineering literature [1, 7, 10].

3.1 Data Description and Preparation

TJSP civil servers selected twelve lawsuit decisions comprising four themes—0929, 1015, 1033, and 1101—each with a designated difficulty level for classification. The difficulty levels are categorized as follows:

- **Easy (E)**: The lawsuit text explicitly mentions the repetitive theme, e.g., "Theme 1033".
- **Medium (M)**: The text contains easily identifiable keywords, though it lacks an explicit mention of the repetitive theme.
- **Hard (H)**: The text neither mentions the repetitive theme explicitly nor contains easily identifiable keywords.

We employed *LegalClass*'s online platform to test each of the four machine learning classifiers—SVM, RL, NB, KNN—individually and in various combinations, resulting in fifteen distinct approaches. Subsequently, we evaluated the probabilities associated with each classification to gauge the confidence level



Classificação

Opções:

○ online ○ batch ● consulta

Opção: consulta

Escolha o arquivo:

acordaos_sem_tema_50_LR_30112022_212312.csv

	numero	id_documento	0929	1015	1033	1039	1101
10	1001010132019826O482	101429657 - 0	0.87	0.01	0.04	0.06	0.02
11	10210465020188260405	82604314 - 0	0.81	0.02	0.08	0.04	0.04
12	10058590220188260405	99212242 - 0	0.73	0.02	0.07	0.14	0.04
13	22649345120208260000	93127415 - 0	0.57	0.04	0.21	0.09	0.09

	0929	1015	1033	1039	1101
count	50.00	50.00	50.00	50.00	50.00
mean	0.65	0.04	0.13	0.09	0.09
std	0.16	0.02	0.07	0.04	0.06
min	0.25	0.00	0.01	0.01	0.01
25%	0.55	0.03	0.08	0.06	0.04
50%	0.62	0.04	0.14	0.09	0.09
75%	0.76	0.05	0.18	0.11	0.11
max	0.98	0.06	0.28	0.18	0.40

	Probabilidade	0929	1015	1033	1039	110
0	#(>=50%)	44	0	0	0	0
1	%(>=50%)	88%	0%	0%	0%	0%
2	#(>=60%)	28	0	0	0	0
3	%(>=60%)	56%	0%	0%	0%	0%
4	#(>=70%)	17	0	0	0	0
5	%(>=70%)	34%	0%	0%	0%	0%
6	#(>=80%)	12	0	0	0	0
7	%(>=80%)	24%	0%	0%	0%	0%
8	#(>=90%)	4	0	0	0	0
9	%(>=90%)	8%	0%	0%	0%	0%

Fig. 3. Query Classification Interface.

of predictions. Each lawsuit text was tested under 180 scenarios comprising individual classifiers and their committees.

3.2 Objective

The experiment aims to answer the research question (RQ): Can *LegalClass* efficiently identify the most effective classification approach for repetitive lawsuit themes?

Our hypothesis (H) posits that *LegalClass*'s probability outputs aid in selecting the best classification approach.

3.3 Variables

Independent Variables:

- Classifiers: Four ML classifiers and their combinations.
- Texts: Twelve lawsuit decisions.

Dependent Variables:

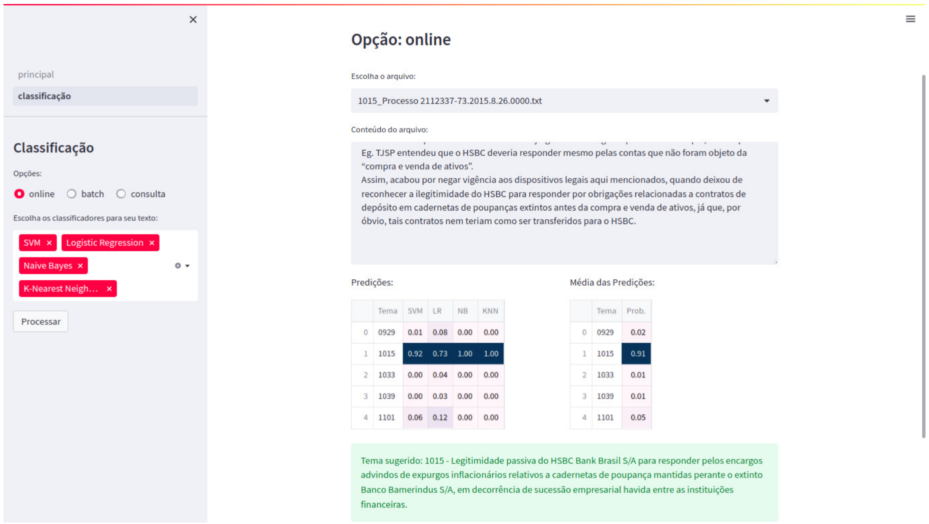
- Response: Classified as right or wrong.

3.4 Results

We present the experimental results obtained through machine learning classifiers on the *LegalClass* platform. As an illustrative example, Fig. 4

showcases the output generated using the "A15" approach, which employs a committee comprising the classifiers SVM, RL, NB, and KNN. The document under consideration is denoted as ”T4,” a text from theme 1015 that is categorized as ’easy’ to classify and carries the process number 2112337-73.2015.8.26.0000.

In this particular instance, all classifiers successfully identified the correct theme. The confidence level in identifying theme 1015 was as follows: 92% for SVM, 73% for LR, for both NB and KNN. When combining the results from all four classifiers into the committee, the overall probability of correctly identifying theme 1015 was 91%.



Classificação

Opções:

● online ○ batch ○ consulta

Escolha os classificadores para seu texto:

Opção: online

Escolha o arquivo:

1015_Processo 2112337-73.2015.8.26.0000.txt

Conteúdo do arquivo:

Eg. TJSP entendeu que o HSBC deveria responder mesmo pelas contas que não foram objeto da "compra e venda de ativos".
Assim, acabou por negar vigência aos dispositivos legais aqui mencionados, quando deixou de reconhecer a ilegitimidade do HSBC para responder por obrigações relacionadas a contratos de depósito em cadernetas de poupanças extintos antes da compra e venda de ativos, já que, por óbvio, tais contratos nem teriam como ser transferidos para o HSBC.

Predições:

	Tema	SVM	LR	NB	KNN
0	0929	0.01	0.08	0.00	0.00
1	1015	0.92	0.73	1.00	1.00
2	1033	0.00	0.04	0.00	0.00
3	1039	0.00	0.03	0.00	0.00
4	1101	0.06	0.12	0.00	0.00

Média das Predições:

	Tema	Prob.
0	0929	0.02
1	1015	0.91
2	1033	0.01
3	1039	0.01
4	1101	0.05

Tema sugerido: 1015 - Legitimidade passiva do HSBC Bank Brasil S/A para responder pelos encargos advindos de expurgos inflacionários relativos a cadernetas de poupança mantidas perante o extinto Banco Bamerindus S/A, em decorrência de sucessão empresarial havida entre as instituições financeiras.

Fig. 4. Example of Classification Output.

Table 2 shows the responses (right (✓) or wrong(✗)) to output the question approach righted or not the theme of the lawsuit decision.

According to Tables 2, incorrect predictions were made only for the texts ”T8” and ”T12.” Text ”T8” is of medium difficulty and is associated with theme 1033, carrying the process number 2080844-73.2018.8.26.0000.

For ”T8,” inaccurate predictions occurred when using the ”A3” and ”A8” approaches. Specifically, the ”A3” approach uses the NB classifier, while ”A8” employs a committee comprising the RL and NB classifiers. The confidence level for the text correctly belonging to theme 1033 was low: 22% for ”A3” and 39% for ”A8.”

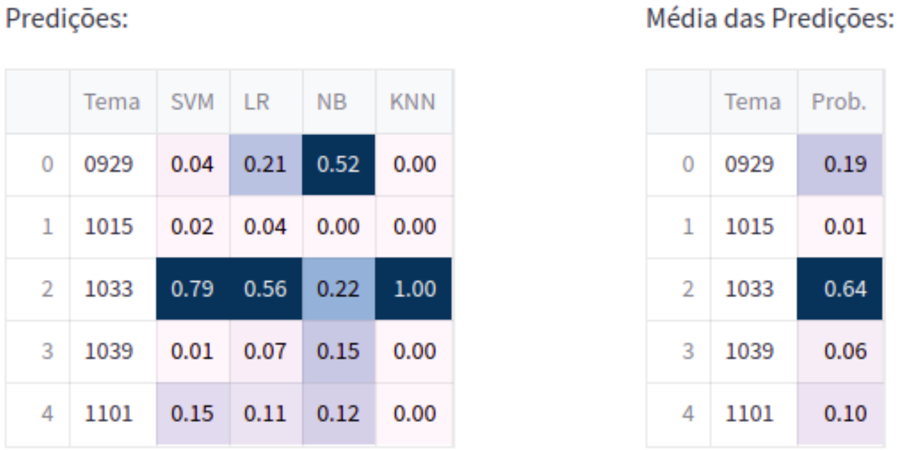
Fig. 5 presents a snippet of the *LegalClass* output when the ”A15” approach was employed, which involves a committee of all four classifiers (SVM, RL, NB, KNN). The first frame of the figure indicates the individual probabilities from

Table 2. Response of Classifiers or Committees

Appr.		**Texts**											
		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
A1	SVM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A2	LR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A3	NB	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
A4	KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A5	SVM, LR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A6	SVM, NB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A7	SVM, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A8	LR, NB	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
A9	LR, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A10	NB, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A11	SVM, LR, NB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A12	SVM, LR, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A13	SVM, NB, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A14	LR, NB, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
A15	SVM, LR, NB, KNN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

each classifier for the text belonging to theme 1033: 79% for SVM, 56% for LR, 22% for NB, and 100% for KNN.

The second frame displays the average probability, considering all four classifiers. In this case, the overall probability of correctly identifying theme 1033 was 64%, as determined by the committee of four classifiers.



Predições:

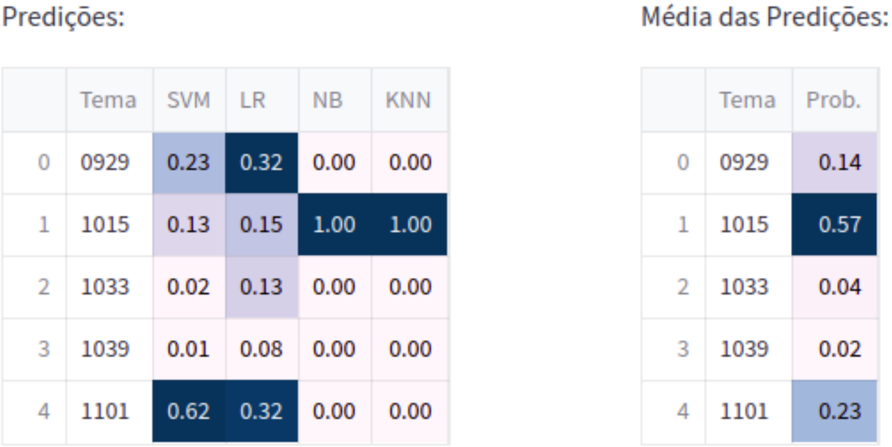
	Tema	SVM	LR	NB	KNN
0	0929	0.04	0.21	0.52	0.00
1	1015	0.02	0.04	0.00	0.00
2	1033	0.79	0.56	0.22	1.00
3	1039	0.01	0.07	0.15	0.00
4	1101	0.15	0.11	0.12	0.00

Média das Predições:

	Tema	Prob.
0	0929	0.19
1	1015	0.01
2	1033	0.64
3	1039	0.06
4	1101	0.10

Fig. 5. Probabilities of Text "T8" by *LegalClass*.

Text ”T12” is categorized as having a high difficulty level for classification and is associated with theme 1011. It carries the process number 2060043-



Predições:

	Tema	SVM	LR	NB	KNN
0	0929	0.23	0.32	0.00	0.00
1	1015	0.13	0.15	1.00	1.00
2	1033	0.02	0.13	0.00	0.00
3	1039	0.01	0.08	0.00	0.00
4	1101	0.62	0.32	0.00	0.00

Média das Predições:

	Tema	Prob.
0	0929	0.14
1	1015	0.57
2	1033	0.04
3	1039	0.02
4	1101	0.23

Fig. 6. Probabilities of Text "T12" by *LegalClass*.

78.2014.8.26.0000. Interestingly, only the ”A1” approach, which utilizes the SVM classifier, correctly identified the text's theme with a confidence level of 62%. All other approaches led to incorrect predictions.

Fig. 6 illustrates a segment of the output generated by *LegalClass* when employing the ”A15” approach, a committee of all four classifiers (SVM, RL, NB, KNN), to categorize ”T12.” The first frame of the figure details the individual probabilities assigned by each classifier for the text being part of theme 1101: 62% by SVM, 32% by LR, 0% by NB, and m 0% by KNN. The second frame displays the average probability, aggregating results from all four classifiers. According to this committee-based approach, the overall likelihood of ”T12” belonging to theme 1101 was a mere 23%.

The tool *LegalClass* facilitated the evaluation of the prediction accuracy and confidence levels for SVM, LR, NB, and KNN classifiers across various text difficulty levels as classified by TJSP civil servers.

Easy Texts: All approaches accurately predicted the themes for texts considered 'easy' (T1, T4, T7, T10). For instance, the likelihood of text T1 belonging to theme 929 exceeded 97% across all approaches, while for text T4 and theme 1015, the probability was above 73

Medium Texts: For 'medium' difficulty texts (T2, T5, T8, T11), all methods were accurate except for text T8. This text yielded incorrect predictions when evaluated using approaches A3 and A8.

Hard Texts: Among the texts classified as 'hard' (T3, T6, T9, T12), only text T12 posed a challenge; accurate prediction was achieved solely through the A1 approach using the SVM classifier.

Theme-specific Insights:

- Texts under theme 929 consistently exhibited the highest probabilities of accurate classification, surpassing 94

- For theme 1015, probabilities remained above 67
- Themes 1033 and 1101 revealed the most variability in prediction confidence, ranging from 22

Research Conclusion: Based on the response and probability data gathered through the *LegalClass* interface, we recommend the Support Vector Machine (SVM) as the most reliable classifier for predicting whether a text belongs to one of the five repetitive themes under study.

4 Conclusion

As judicial workloads grow, automation and modern techniques are increasingly essential. One challenge in Brazil's judicial system is efficiently classifying lawsuits into predefined repetitive themes. This paper presented *LegalClass*, a web-based tool designed to tackle this issue by employing machine learning algorithms for text classification.

Our tool allows for categorising individual or multiple legal texts and provides theme suggestions and associated probabilities. Though initially equipped with four traditional machine learning algorithms, the tool's architecture permits the straightforward inclusion of additional algorithms and themes.

Using a case study, we demonstrated the tool's efficacy by employing it on real-world data categorized by civil servers at TJSP. The SVM algorithm emerged as the most effective classifier, correctly categorizing all provided texts and offering an intuitive, user-friendly interface.

Automated text classification holds significant promise for enhancing judicial efficiency, but further research and development are required to refine this technology. Future advancements in this domain can yield more precise and efficient systems, aiding decision-makers in navigating extensive data sets more effectively.

References

1. Basili, V.R., Rombach, D., Kitchenham, K.S.B., Selby, D., Pfahl, R.W.: Empirical Software Engineering Issues. Springer Berlin/Heidelberg (2007)
2. Darasay, B.V.: Nearest neighbor pattern classification techniques, los alamitos (1991)
3. Del Mar-Raave, J.R., Bahşi, H., Mrsic, L., Hausknecht, K.: A machine learning-based forensic tool for image classification - a design science approach. Forensic Science International: Digital Investigation **38**, 301265 (2021)
4. Hearst, M., Dumais, S., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their Applications **13**(4), 18–28 (1998)
5. andFatma Hilal Yagin, I.P., Arslan, A.K., Colak, C.: An interactive web tool for classification problems based on machine learning algorithms using java programming language: Data classification software. In: International Symposium on Multidisciplinary Studies and Innovative Technologies. pp. 1–7 (2019)
6. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. John Wiley & Sons (2013)

7. Jedlitschka, A., Pfahl, D.: Reporting guidelines for controlled experiments in software engineering. In: Int. Symposium on Empirical Software Engineering. pp. 95–104 (2005)
8. Losada, D.E., Azzopardi, L.: Assessing multivariate bernoulli models for information retrieval. ACM Transactions on Information Systems (TOIS) **26**(3), 1–46 (2008)
9. de Justiça Departamento de Pesquisas Judiciárias, C.N.: Justiça em números 2021. Justiça em números 2021 (2021 [Online])
10. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering. Springer Science & Business Media (2012)
11. Zhang, C., Bi, J., Xu, S., Ramentol, E., Fan, G., Qiao, B., Fujita, H.: Multi-imbalance: An open-source software for multi-class imbalance learning. Knowledge-Based Systems **174**, 137–143 (2019)