São Paulo, Brazil
November 4-5, 2019

The Eleventh International Conference on
FORENSIC COMPUTER SCIENCE and CYBER LAW

www.ICoFCS.org

DOI: 10.5769/C2019001 or http://dx.doi.org/10.5769/C2019001

Descoberta de termos que caracterizam peças jurídicas

Davi Alves Bezerra[1], Pedro H. G. Inazawa[2], Roberta Zumblik[3], Teófilo E. de Campos[4], Nilton Correia da Silva[5], Fabrício A. Braz[6], Fabiano Hartmann Peixoto[7].

(1) Faculdade do Gama, Universidade de Brasília, Email: davialvb@gmail.com
(2) Faculdade do Gama, Universidade de Brasília, Email: pedro.inazawa@gmail.com
(3) Faculdade de Direito, Universidade de Brasília, Email: robertazumblickms@gmail.com
(4) Depto. de Ciência da Computação (CIC), Universidade de Brasília, Email: t.decampos@oxfordalumni.org
(5) Faculdade do Gama, Universidade de Brasília, Email: niltoncs@unb.br
(6) Faculdade do Gama, Universidade de Brasília, Email: fabraz@unb.br
(7) Faculdade de Direito, Universidade de Brasília, Email: fabiano_hp@hotmail.com

Abstract: The Brazilian Court System is currently the biggest judiciary system in the world, and has millions of scanned processes. However, it suffers from low efficiency problems due to low investment in new technologies. The usage of AI (Artificial Intelligence) is a suitable solution, since it can automate tasks that could, in recent past, only be done by humans. To facilitate the analysis of data, we propose to automatically discover the relevance of keywords. For that, we evaluate two classifiers (Naive Bayes and K-Nearest Neighbors) using Chi-squared statistical analysis and Term Frequency-Inverse Document Frequency count (TF-IDF) for feature ranking. Our experiments were performed on a dataset of 5 classes of judiciary pieces obtained from Brazil's Supreme Court (STF). Finally, a feature vector of the best ranked words is compared to the legal jargon of each selected piece.

Key words: Brazilian Court System, Artificial Intelligence, Naive Bayes, K-Nearest Neighbors, TF-IDF, Chi-squared.

I. Introdução

O Brasil possui um dos maiores sistemas jurídicos do mundo em volume de processos [4]. O Poder judiciário é referência entre os três poderes em termos de transparência e levantamento de dados, e tem - desde 2004 - uma pesquisa denominada Justiça em Números, que permite o conhecimento dos dados e gargalos processuais e serve como instrumento de planejamento e gestão. O relatório divulgado em 2018, atualizado com os dados do ano anterior, aponta que 2017 foi o ano com o menor crescimento de acervo - mas que ainda assim cresceu em 0,3%, totalizando 80,1 milhões de processos aguardando uma solução definitiva [4]. O grande volume de processos judiciais no Brasil é ao mesmo tempo um problema e uma oportunidade para o desenvolvimento de novas soluções e tecnologias. Já existem diversas tecnologias baseadas no estudo de grandes massas de dados e automatização que poderiam auxiliar na

realização de tarefas repetitivas, tramitação de processos e também servirem de apoio a decisões. Essas tecnologias não somente ajudariam os tribunais, como também todo o Poder Público e demais seguimentos que tem uma grande quantidade de processos, como seguradoras, empresas de telecom, bancos e e-commerces.

I.I Sobre o trabalho

O presente trabalho apresenta a comparação entre duas técnicas de avaliação de características para Processamento de Linguagem Natual (PLN) sobre um dataset de 5 tipos de peças jurídicas, e demonstra o resultado de F1-Score para dois métodos classificadores muito utilizados em aprendizado de máquina. Por fim, o trabalho compara o jargão conhecido de cada uma das peças com as palavras que os algoritmos consideraram mais relevantes, promovendo uma reflexão sobre uma espécie de assinatura de peças.

2. Metodologia

O método empregado para seleção de características no universo de documentos estudados consiste em ranquear as potenciais palavras mais importantes para as classes avaliadas a partir dos procedimentos descritos a seguir.

2.1 Pré-Processamento

O texto é extraído usando um OCR [11] executado em documentos em formato PDF. Esse conjunto cru de dados obtido é submetido a um processamento necessário para adequá-lo às análises posteriores. Isso foi feito por meio de analise léxica no conjunto de caracteres desejados para a produção de símbolos léxicos segundo as seguintes etapas, visando a eliminação de palavras com pouca ou nenhuma informação, e realce de outras com interesse para o domínio:

1) Remoção de caracteres especiais, tais como #, @, $, e afins;
2) Remoção de termos alfanuméricos, isto é, com números e letras mesclados;
3) Transformação dos termos que sejam e-mails ou links nos símbolos léxicos “EMAIL” e “LINK”;
4) Transformação dos termos referentes a números de leis e artigos nos símbolos léxicos “LEI_X” e “ARTIGO_X”, onde o X representa a respectiva lei ou artigo citado;

2.2 Criação do vetor de descritores

Neste trabalho, avaliamos dois métodos para se determinar a relevância de cada termo do vocabulário: teste estatístico x^2, e Teste de $TFIDF$ [5], [6]. Com base nos resultados desses métodos citados, foi gerado um vocabulário de palavras características para cada peça em questão com os termos ranqueados.

Chi-quadrado: O teste $x^2(t,c)$ mede a dependência entre duas distribuições t e c. A fórmula que descreve o teste é a seguinte:

$$x^2(t,c) = \frac{N(AD - CB^2)}{(A+C)(B+D)(A+B)(C+D)}, \qquad (1)$$

Sendo que A é a quantidade de vezes que t e c co-ocorrem, B é a quantidade de vezes que t ocorre sem c, C é o número de vezes que c ocorre sem t, D é o número de vezes que nem c nem t ocorrem e N é o número total de documentos. Esse valor é naturalmente mais próximo de zero conforme as distribuições sejam mais independentes, e vice-versa. Para este trabalho, foi feita uma avaliação baseada no valor individual de cada categoria avaliada, e no valor conjunto somado dos testes.

Frequência do termo - frequência inversa do documento (ou TF-IDF, do inglês): Essa é uma medida estatística utilizada no estudo da teoria da informação que mede o quão importante é uma

palavra para classes de documentos com base na sua frequência nos corpos textuais, e na quantidade de documentos em que se encontra presente. É composta de duas etapas: a Frequência normalizada de Termos (do inglês, Term Frequency ou TF), que mensura quão frequente são os termos analisados em um documento, e a Frequência Inversa de Documentos, que quantifica o quão importante um termo é, dada a sua raridade no conjunto de documentos avaliados. O resultado final é multiplicação de ambas as quantidades, conforme mostrado na equação abaixo:

$$TFIDF(t) = TF(t) \cdot IDF(t), \qquad (2)$$

sendo que $TF(t) = \frac{q_t}{Q_D}$, $IDF(t) = \log_e(\frac{N_D}{n_t})$, q_t é a quantidade de vezes que o termo t aparece em um documento, Q_D é a quantidade total de termos em um documento, N_D é a quantidade total de documentos, n_t é a quantidade de documentos com o termo t presente.

2.3 Classificação

Para se verificar o desempenho dos vetores de termos criados conforme descrito na Seção foram avaliados dois classificadores, o Naive Bayes [9] e o de K vizinhos mais próximos (KNN) [7].
O classificador Naive Bayes foi selecionado para este teste por ser um classificador baseado num modelo generativo dos dados.
Esse modelo é muito simples e por isso ele é extremamente rápido, com complexidade linear na dimensionalidade dos dados. Apesar dessa simplicidade, resultados tendem a ser bem satisfatórios em problemas com alta dimensionalidade, como os problemas de classificação de texto.
Em contrapartida, o KNN foi selecionado porque, em princípio, em bases de treinamento com grandes conjuntos de amostras altamente representativas da distribuição dos dados e sem *outliers*, o classificador de 1NN (KNN com K=1 vizinho) levaria a resultados ótimos. O uso de valores de K>1 aumenta a robustez contra *outliers*. Nos nossos experimentos, utilizamos K=5, que coincide com o número de classes que a base de peças jurídicas possui. A fase de treinamento desse classificador é trivial, pois consiste em simplesmente armazenar todos os vetores. Porém, na fase de teste, cada amostra deve ser comparada com todas as amostras de treinamento e os resultados devem ser ordenados. Há vários algoritmos aproximados que tornam inferência mais rápida, por exemplo [8].
Os resultados foram avaliados usando Precisão, Recall e o F1-Score [2].

3. Conjunto de dados utilizado

Utilizamos o conjunto de dados do Projeto Victor, publicado em [3]. Esses dados têm origem em processos que ingressam o Supremo Tribunal Federal (STF) do Brasil. Os documentos obtidos vieram em grandes volumes, isto é, arquivos sem divisões e onde não havia separação entre peças. Uma equipe de especialistas em direito foi então acionada para fazer o rótulo desse material, separando-o em 5 tipos de peças principais e um grupo ”Outro” (esse agrupa todas as demais peças). Os tipos de peças foram: “Acórdão”, “Recurso Extraordinário (RE)”, “Agravo de Recurso Extraordinário (ARE)”, “Decisão de admissibilidade” e “Sentença”.
A divisão do dataset foi 70% dos registros para treino, 10% para teste e 20% para validação.
Na Figura 2, é apresentada a quantidade de amostras para cada um dos subconjuntos. Despachos e RE foram os tipos de documentos que menos tiveram amostras, presentes na Figura 1.
Para este trabalho, foi considerado o conjunto de Treino e de Teste do dataset apresentado, isto é, todos os algoritmos apresentados foram treinados com o conjunto de Treino e os resultados expostos foram feitos a partir do conjunto de Teste. Não utilizamos o conjunto de validação por não se ter a necessidade de validações iniciais durante a aprendizagem de nossos modelos.

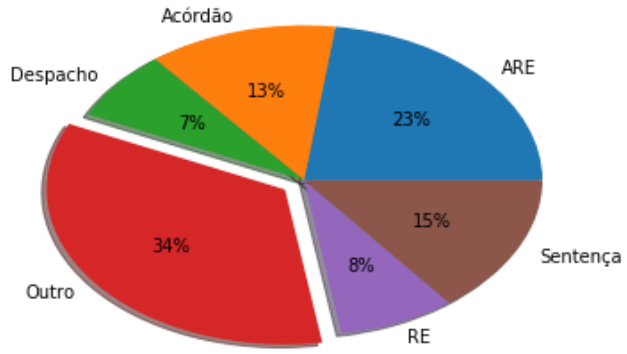


Figura 1 – Total de peças no conjunto de dados, reproduzida de [3].

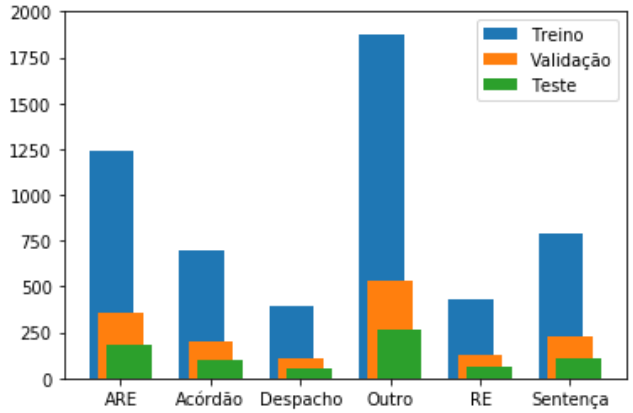


Figura 2 – Divisão da base de dados, reproduzida de [3].

4. Resultados e discussão

A seção 4.1 comenta o resultado dos classificadores e do método de seleção de características. Seção 4.2 a segunda comenta acerca das características obtidas, e como elas se relacionam ao domínio estudado.

4.1 Modelos testados e seleção de termos mais relevantes

A Figura 3 ilustra a performance média dos F1-Scores dos classificadores testados com base em ambos os esquemas de pesos propostos.

Com relação a vetorização de palavras, o x^2 foi o método que melhor classificou para uma menor quantidade de palavras exploradas em ambos os cenários. O $TFIDF$ acabou tendo uma performance inferior, que atingiu em torno de 80% no melhor cenário obtido. Isso está relacionado ao foco de cada um dos algoritmos: enquanto o TF-IDF privilegia palavras mais raras dentro de uma mesma classe, o Chi-quadrado avalia dependências entre as distribuições das palavras e das classes, se tornando uma métrica mais forte de diferenciação [1], [10]. Ao se tratar de seleção de características, é possível visualizar e confirmar que o x^2 é também uma excelente abordagem para uma área muito especifica.
Com relação ao modelo de classificação utilizado, o KNN apresentou dois resultados diferentes. Com x^2, a classificação inicial com um espaço de características (palavras) pequeno foi de alto F1-score, e conforme introduzidas mais palavras, o resultado piorou. Já com TF-IDF, o resultado começou num patamar baixo e, conforme as palavras foram sendo inseridas, o resultado melhorou (não atingindo, no entanto, o mesmo resultado que a abordagem usando x^2). O modelo Naive Bayes em ambas as abordagens acabou reagindo de maneira semelhante: conforme o número de características aumentou, também diminuiu o F1-Score. O resultado com TF-IDF, no entanto, obteve pontuação inferior ao com x^2.
A Figura 4 é um gráfico de Quantidade de Informação por Quantidade de termos presentes do vocabulário. A quantidade de informação no caso deste gráfico foi obtida a partir da normalização dos valores da melhor vetorização utilizada (x^2) Equação 1. O gráfico demonstra que pouco menos de 15% das palavras já possuem algo em torno de 85% de toda a informação presente no vocabulário de todas as peças. Dessa forma, é possível gerar um conjunto de palavras bem mais expressiva para a caracterização das classes avaliadas.

4.2 Palavras selecionadas e seu relacionamento com o domínio

Com base na informação presente nos gráficos da Figura 3 e Figura 4, foi possível estabelecer um conjunto característico das palavras que contém mais informação para cada classe, estabelecendo assim um vocabulário (ou jargão) mais utilizadas dentro do universo das classes avaliadas. A Figura 5 ilustra esse resultado para as 30

melhores (mais influentes) palavras de cada classe.
O resultado inicial das palavras mais influentes de cada classe pode ser analisado sob diversas perspectivas e, consequentemente, demonstram importantes características. Nesse estudo serão apresentados alguns desses aspectos vistos à luz da experiência jurídica na análise de peças, isto é, algumas questões que podem sugerir a relação da importância das palavras e seu vínculo com o domínio.

Examinando-se as palavras encontradas e selecionadas por classe de documento, pode-se dizer que as ordens de palavras mais relevantes de fato fazem sentido, isto é, relacionam-se intimamente com o domínio a qual dizem respeito. Nesse sentido, para citar alguns, os vocábulos

1) “reflexa”, “infraconstitucional”, “constitucionais'”, “violação”, “repercussão”, que aparecem na Decisão de Admissibilidade;
2) “artigo”, “102”, “iii”, “alínea”, que aparecem no Recurso Extraordinário;
3) “sentença”, “voto”, “relator”, que aparecem no Acórdão, etc; são termos utilizados com frequência na construção das peças destacadas, tendo em vista que servem para caracterizar aquela classe determinada.

Há que se ressaltar o aparecimento notório de alguns vieses, como por exemplo a palavra “benefícios” que aparece na identificação de várias peças como ARE, Sentença, RE e Acórdão, que diz respeito a um tema que muito se apresenta nas peças, qual seja, as questões que envolvem benefícios previdenciários. Além disso, há palavras que identificam artigos do Código de Processo Civil de 1973 (o qual esteve vigente até março de 2016), como por exemplo o “Artigo 544” que aparece no ARE, e que servem para identificar classes atualmente, mas que no futuro não serão mais utilizadas, tendo em vista a entrada em vigor no Novo Código de Processo Civil de 2015. Esses vieses possivelmente aparecem, pois o dataset dessa fase do Projeto ainda não é genérico o suficiente, além de representarem dados dos temas mais recorrentes no STF nos últimos dois anos (2016-2018).

Observa-se, por fim, a ocorrência reiterada de palavras que são demasiado genéricas (como por exemplo o vocábulo “senhor” encontrado no ARE; “documento” que aparece no RE) cujo sentido e utilização pode se dar em contextos variados que não necessariamente relacionam-se especificamente com determinada peça.

5. Conclusão e Considerações finais

A proposta de utilização de técnicas clássicas de Processamento de Linguagem Natural (PLN) juntamente com Aprendizado de Máquina (AM) conforme apresentado neste trabalho pode gerar impactos positivos na eficiência de toda gestão pública brasileira, bem como em diversos setores que possuem grandes quantidades de processos jurídicos em andamento. Os resultados obtidos revelaram que apenas cerca de 15% das palavras avaliadas já representam em torno de 85% de toda a informação dos domínios, as técnicas de classificação apresentadas podem auxiliar metodologias de AM que busquem usar grandes massas de dados jurídicos, focando apenas nas características mais relevantes do texto.
Como trabalhos futuros, pretende-se aplicar as técnicas apresentadas e algumas outras relacionadas a quantidade de informação em datasets jurídicos maiores, para verificar a qualidade das características obtidas. Outro trabalho futuro será a avaliação de metodologias de seleção de características com o uso de combinação de palavras.

Referências

[1] S. Brindha, Prabha. The comparison of term based methods using text mining. International Journal of Computer Science and Mobile Computing, 5:112–116, 09 2016.

[2] L. Cuadros-Rodríguez, E. Pérez-Castaño, and C. Ruiz-Samblás. Quality performance metrics in multivariate classification methods for qualitative analysis. TrAC Trends in Analytical Chemistry, 80, 04 2016.

[3] N. C. da Silva, F. A. Braz, T. E. de Campos, D. Gusmao, F. Chaves, D. Mendes, D. Bezerra, G. Ziegler, L. Horinouchi, M. Ferreira, G. Carvalho, R. V. C. Fernandes, F. H. Peixoto, M. S. M. Filho, B. P. Sukiennik, L. S. Rosa, R. Z. M. Silva, and T. A. Junquilho. 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 29-30 2018. http://icofcs.org.

[4] L. Fariello. CNJ apresenta justiça em números 2018, com dados dos 90 tribunais. Online, 28 August 2018. http://www.cnj.jus.br/noticias/cnj/87512-cnj-apresenta-justica-em-numeros-2018-com-dados-dos-90-tribunais.

[5] G. Forman. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. (JMLR), 3:1289–1305, Mar. 2003.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. J. Mach. Learn. Res. (JMLR), 3:1157–1182, Mar. 2003.

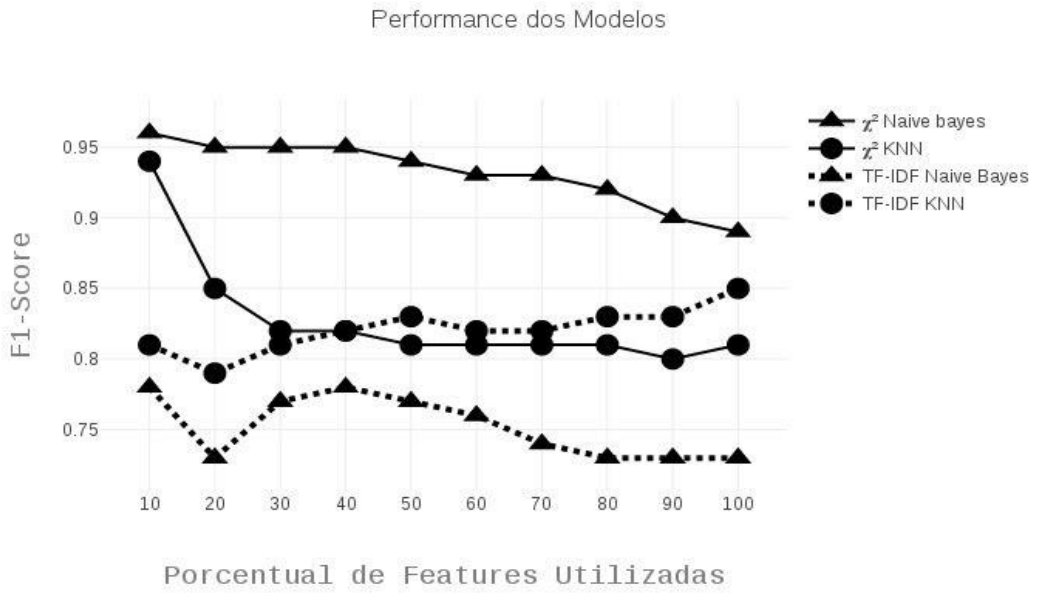
[7] S. Kr. Srivasatava, R. Kumari, and S. Kr. Singh. An ensemble based NLP feature assessment in binary classification. In International Conference on Computing, Communication and Automation (ICCCA), pages 345–349, 05 2017.

[8] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. VISAPP (1), 2(331-340):2, 2009.

[9] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In In Proceedings of the Twentieth International Conference on Machine Learning (ICML), pages 616–623, 2003.

[10] M. Rogati and Y. Yang. High-performing feature selection for text classification. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM'02, pages 659–661, New York, NY, USA, 2002. ACM.

[11] R. Smith. An overview of the Tesseract OCR engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR), volume 2, pages 629–633. IEEE, 2007.



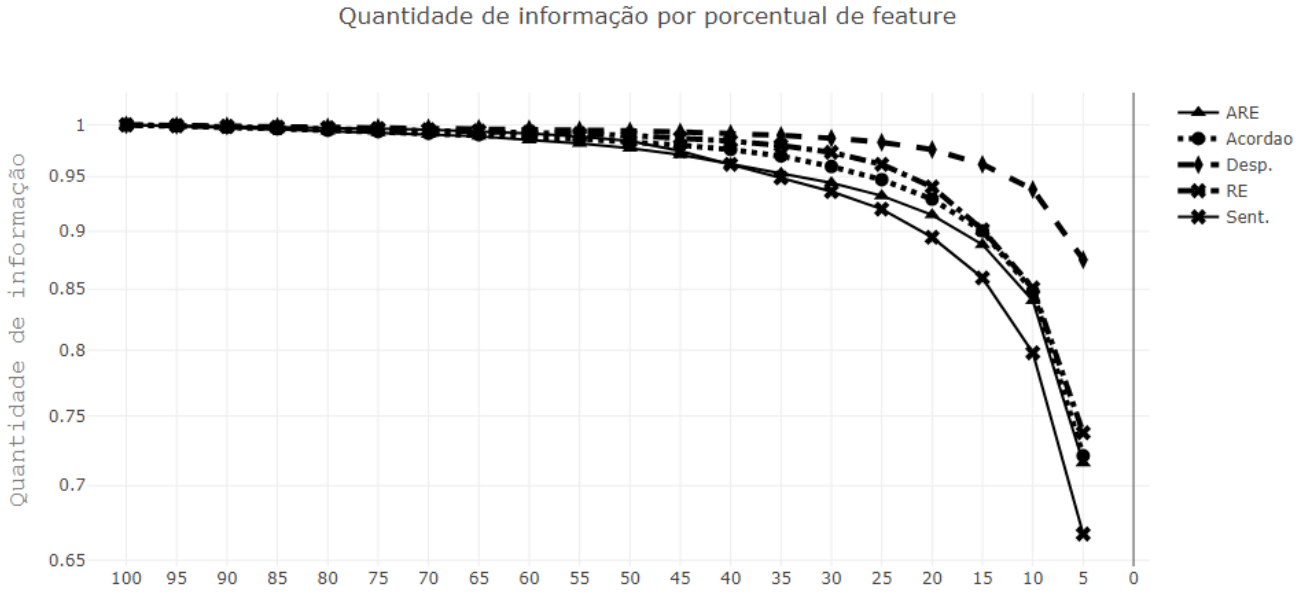


Figure 4 - Quantidade de informação mantida em cada tipo de peça, dado o percentual de termos mantidos no vocabulário.

Figure 5 – Nuvens de palavras com os 30 elementos mais relevantes de cada classe. Os gráficos foram organizados de modo que as palavras em maior tamanho são as de impacto superior na classificação. Dessa forma, quanto maior o tamanho da palavra, maior a importância na classificação.