

Sistema de Suporte à Decisão Inteligente em Processos de Herança:

uma Pesquisa na Procuradoria Geral do Estado de Pernambuco

Intelligent Decision Support System in Inheritance Processes:

a Research in the General Attorney Office of the State of Pernambuco

Resumo — O setor jurídico público brasileiro é tipicamente marcado pela razão desproporcional entre o número de processos e a quantidade disponível de funcionários especializados. No entanto, muitos desses processos são de menor importância para o Estado, o que resulta em sobrecarga de funcionários altamente especializados e caros ao orçamento público com atividades que poderiam ser desempenhadas automaticamente ou redistribuídas para profissionais de nível hierárquico inferior e menos onerosos. O presente trabalho propõe uma abordagem de suporte à decisão em análise de processos de herança em uma procuradoria estadual. O objetivo é desenvolver um sistema de coleta, classificação, análise e ranqueamento de processos usando de mineração de texto e aprendizagem de máquina a fim de automatizar o procedimento de análise de documentos de maneira semântica, e distribuir objetivamente os processos, segundo a sua importância monetária, de acordo com o nível hierárquico dos profissionais. O sistema dispõe de uma análise semântica calibrável, que leva em conta a experiência do usuário durante o ranqueamento dos documentos, e que pode ser facilmente adaptável para outras atividades de análise em procuradorias.

Palavras Chave - mineração de texto, aprendizagem de máquina, procuradoria, análise automática de processos.

Abstract — The Brazilian public legal sector is typically marked by the disproportionate ratio between the number of cases and the available number of specialized employees. However, many of these processes have minor importance for the State, which promotes overload for highly specialized and expensive employees with non-important activities that could be performed automatically or redistributed to lower level professionals. The present work proposes a decision support approach for analyzing inheritance processes in a state attorney's office. The objective is to develop a system for collecting, classifying, analyzing and ranking processes using text mining and machine learning in order to automate the document analysis procedure in a semantic

manner. The put forward approach offers a novel way for objectively distribute processes according to their monetary importance and the hierarchical level of the involved professionals. The approach also proposes a tunable semantic analysis, which takes into account the user experience for the automatic document ranking, which can be easily adapted for other analytical activities in attorney's offices.

Keywords - text mining, machine learning, attorney, automatic process analysis.

I. INTRODUÇÃO

A governança de uma região exige que o estado seja responsável pela manutenção do bem público, nesse contexto, para arrecadação da quantia devida ao funcionamento do governo, são realizadas atividades de cobrança de impostos e multas, assim como a fiscalização sobre o devido pagamento dessas contribuições por parte dos cidadãos. Como braço jurídico do Poder Executivo a Procuradoria tem por finalidade auxiliar o Poder Executivo a garantir a integridade dos direitos do Estado [1]. Parte das atividades atribuídas às procuradorias está na cobrança, judicialização e acompanhamento de processos relativos a dívidas tributárias não pagas ou indevidamente pagas ao estado.

Um dos tipos de dívidas públicas, de responsabilidade de uma procuradoria estadual no Brasil, são as decorrentes de impostos de herança. Ao herdar bens ou valores monetários, os herdeiros devem pagar o imposto proporcional ao valor da herança [2], no entanto, é costumeiro que alguns imóveis sejam registrados no plano de partilha com valores abaixo do de mercado, assim como omissão de bens durante o cálculo dos impostos. Cabe à procuradoria designar um núcleo de funcionários para realização das atividades de fiscalização e análise de processos de sucessão a fim de averiguar

inconsistências e tomar as medidas cabíveis para garantir o interesse público.

O trabalho numa procuradoria é marcado pela necessidade de analisar uma grande quantidade de documentos, que chegam diariamente, confrontado por uma pequena quantidade de procuradores, altamente qualificados e de elevado custo para o estado. Esse contraste consequentemente resulta em sobrecarga e acúmulo de processos. No entanto, muitos desses processos são de pouca importância monetária para o estado, uma vez que o falecido pode ter deixado nenhum ou apenas uma pequena quantidade de bens. Contudo, mesmo sendo processos de menor importância para o estado, e por mais que possa haver erros no cálculo das taxas de imposto a serem pagas, o custo por hora trabalhada do procurador responsável por esse processo é maior que o imposto a ser cobrado.

Por outro lado, o direito administrativo do estado brasileiro segue o princípio da indisponibilidade do interesse público [3], o que obriga que todos os processos sejam devidamente analisados e corretamente cobrados pelo estado.

Atualmente, técnicas de mineração de texto [4], suportadas por algoritmos de aprendizagem de máquina [5], já vêm sendo utilizadas em diversos setores, inclusive no meio jurídico [6], a fim de automatizar atividades que necessitem de inteligência humana. Essas técnicas de mineração de texto e aprendizagem de máquina podem ser usadas para realização de uma série de tarefas relacionadas a documentos [7], indo desde a conversão de arquivos de imagem em texto [8]; pré-processamento de texto, o que inclui limpeza de acentos, pontuação, espaços vazios, erros de conversão, etc.; identificação de termos importantes; análise de sentimentos; classificação de texto, entre outras. Outra vertente a considerar são os sistemas de apoio à decisão [9][10], que auxiliam no processo de análise e aceleram o tempo de decisão em processos que necessitem de inteligência humana. Esses sistemas de apoio à decisão podem ser baseados em regras, ou seja, na experiência do especialista, ou inteligentes, usando de técnicas estatísticas para realização de aprendizagem de máquina.

Nessa perspectiva, técnicas de mineração de texto podem ser utilizadas com a finalidade de extrair informações importantes de documentos, além de automatizar parte do procedimento de análise, auxiliando a tomada de decisão e, consequentemente, minimizando o tempo gasto pelo especialista. Dessa forma, mineração de texto e sistemas de suporte à decisão podem ser ferramentas chave para solução de problemas relativos à sobrecarga de trabalho e lentidão do serviço público.

Nesse contexto, o presente trabalho propõe um fluxo de trabalho baseado em mineração de texto e aprendizagem de máquinas a fim de abranger com plenitude todos os passos necessários para automação, de maneira inteligente, da análise de processos de herança em uma procuradoria estadual. O principal objetivo é que o sistema proposto seja capaz de ranquear processos e auxiliar no processo decisório de divisão e atribuição de trabalho entre profissionais de diferentes níveis de qualificação. Sendo assim, os processos podem ser distribuídos, segundo a sua quantidade e relevância, entre procuradores (altamente qualificados), analistas (qualificados)

e profissionais em formação, como analistas juniores ou estagiários. Como objetivos secundários, busca-se construir uma ferramenta hábil a auxiliar no processo de análise de processos de herança, para acelerar a atividade humana e, consequentemente, a eficiência do estado.

II. METODOLOGIA

A presente abordagem foi elaborada com metodologia CRISP-DM [11] e propõe a realização de 7 etapas, que vão desde a coleta de documentos até o ranqueamento de processos, como mostrado no diagrama da Fig. 1. Por sua vez, cada uma dessas etapas, que serão melhores discutidas a seguir, pode ser constituída de procedimentos baseados em regras, análises semânticas e aprendizagem de máquina.

A. Coleta de Documentos nas Pastas Virtuais

Esse processo pode ser realizado por meio de um robô, que realiza a coleta de todos os documentos disponíveis na pasta de cada processo, disponibilizados no ambiente virtual do governo. Esses documentos podem ser armazenados em um banco local, do qual serão extraídos, durante a realização das etapas seguintes.

B. Conversão de Documentos em Texto

Foi estabelecido que todos os documentos disponibilizados para o sistema necessitam estar em formato PDF. No entanto, alguns dos arquivos disponíveis nas pastas digitais já foram criados em PDF, enquanto outros são imagens digitalizadas e armazenadas neste formato. Se um dado documento está em formato PDF nativo, seus textos podem ser extraídos de maneira automática e rápida por bibliotecas especializadas.

Caso o documento tenha sido digitalizado, a conversão em texto deve passar pelo procedimento de reconhecimento ótico de caracteres (OCR), que, por sua vez, é um processo computacionalmente custoso, lento e suscetível a erros. Para realização de OCR nessa abordagem, foi escolhida a ferramenta Tesseract OCR [8], por ser gratuita, de boa qualidade e amplamente utilizada na academia.

Além da dificuldade de processamento, antes que os textos sejam reconhecidos por meio de uma ferramenta de OCR, se faz necessário que as imagens sejam pré-processadas para otimizar a sua conversão em texto. Um pré-processamento inapropriado pode resultar em textos corrompidos e ilegíveis. O pré-processamento de imagens aqui utilizado consiste em 3 etapas seguidas: eliminação dos canais vermelho e azul (1); conversão em escala de cinza (2) e conversão em preto e branco (3), que pode ser feito por meio de truncagem binária em 50% da escala ou truncagem por binarização Otsu [12].

A fim de aumentar a velocidade de conversão, todos os documentos disponibilizados passam pela retirada de texto em PDFs, caso a quantidade de caracteres obtidos por página seja menor que um parâmetro estabelecido, entende-se que o documento é do tipo digitalizado e deve passar também pelo processo de OCR. Através de uma pesquisa empírica, esse parâmetro foi configurado em 500 palavras.



Figure 1. Fluxograma do processo

C. Pré-processamento de Texto

Uma vez convertidos todos os documentos em texto, esses necessitam de um processo de limpeza e avaliação de qualidade antes de seguirem adiante. Esse procedimento consiste em 4 etapas: retirada de acentos gráficos, pontuação, parágrafos e espaços duplos (1); transformação de todas as letras em minúsculas (2); remoção de palavras de pouca importância, chamadas de *stop words*, que para essa aplicação são: artigos, preposições e conjunções (3); e, por último, avaliação idiomática, caso o texto não esteja em um idioma aceitável, consequência de falhas na conversão de imagem para texto, ele deve ser removido (4).

D. Classificação Inteligente de Documentos

Os processos inseridos no sistema são compostos de uma grande quantidade de documentos. No entanto, apenas os documentos que contém os bens são necessários para automação da análise. Com o intuito de separar os documentos importantes que prosseguirão para a próxima etapa, é proposta a utilização de um algoritmo de aprendizagem de máquina para realização de classificação de uma única classe. O algoritmo utilizado para isso foi o XGboost com a parametrização padrão[13] esse algoritmo foi escolhido devido a sua excelente resposta para classificação[14]. Nessa abordagem, o classificador tem por objetivo ler um dado documento e informar se ele é ou não um documento importante.

Para isso os documentos convertidos precisam passar por um procedimento de *stemming* [6], no qual todas as palavras são reduzidas ao seu tronco, o objetivo é diminuir a quantidade de palavras diferentes por documentos. Após isso, os textos passam por um algoritmo de vetorização com saco de palavras - do inglês *Bag of Words* (BoW) [6] -, onde cada documento passa a ser representado pela quantidade de cada palavra que possui e que compõem o BoW.

Como se trata de classificação uma única classe, o BoW é composto apenas de palavras encontradas nos documentos rotulados como importantes. O vetor de cada texto é usado como entrada para o classificador. Uma vez treinado, o classificador pode ser utilizado corriqueiramente no fluxo de trabalho.

Durante essa etapa, caso para um dado processo sejam identificados mais de um documento importante, é escolhido o documento com maior número de páginas e os demais são descartados.

E. Pontuação Semântica de Processos

Com ajuda do especialista foram identificados 95 termos chave, que foram pontuados de -1 a 20. Essa pontuação leva em conta dois critérios, o primeiro deles é um critério objetivo, o qual referencia o valor dos bens associados, a título de ilustração, a palavra 'motocicleta' vale 2 pontos, enquanto que a palavra 'apartamento' vale 6 pontos.

O segundo critério é subjetivo e leva em conta o grau de fortuna associado ao bem. Esse critério faz com que palavras que simbolizam abundância sejam maior pontuadas, como 'porsche', 'lexus' ou 'iate', que pontuam 15. Um apartamento pode tipicamente ser de maior valor monetário que um veículo de luxo, mas a posse de um veículo desse porte é um forte indicio de riqueza. A presença desse segundo critério dá uma ênfase semântica à análise de processos, de maneira que o especialista possa adaptar o procedimento de busca segundo a sua experiência profissional e entendimento da cultura regional.

Nesse contexto, nomes de bairros nobres podem ser utilizados como termos chave e pontuados de maneira subjetiva à experiência do analista, enquanto que a presença de bens genéricos como "casa", "veículo", "terreno" e "apartamento" são pontuados objetivamente de acordo com o valor monetário médio.

Além da pontuação positiva, alguns termos chave foram avaliados com pontuação neutra (valor 0), como "não deixou bens" e "inventário negativo". Esses termos ajudam a identificar documentos que foram devidamente selecionados, mas que os processos não têm pontuação.

Outros termos chave receberam pontuação negativa (valor -1), como "abertura de inventário" e "testamento". Esses termos normalmente indicam que os documentos disponíveis não correspondem ao processo completo. De forma que processos com pontuação final negativa necessitam de uma verificação humana mais criteriosa quanto à disponibilidade dos documentos.

Um BoW com todos os termos chaves é utilizado para extrair as características dos documentos selecionados pelo classificador. De forma que são encontradas quais palavras chaves e quantas vezes cada uma delas aparece nos processos. Essa informação, junto com a tabela de pontos por termos chave, é utilizada para calcular a pontuação total de cada processo.

F. Classificação Condicional

Uma vez pontuados, os processos podem ser classificados em: Analisáveis, Duvidosos ou Incompletos. Os processos Analisáveis são aqueles com pontuação positiva ou nula, mas que no qual tenha sido encontrado ao menos 1 termo chave. Os duvidosos são os de pontuação nula nos quais não foram encontrados quaisquer termos chave, esses processos necessitam de uma análise humana de média complexidade para se avaliar o motivo da ausência desses termos. Os processos classificados como Incompletos são os de pontuação negativa e necessitam de trabalho humano de média complexidade, a fim de encontrar nos bancos de dados disponíveis os documentos que disponham de informações de bens, uma vez que estavam incompletos nos arquivos

disponibilizados para o sistemas, e devem ser realimentados no início do processo.

G. Ranqueamento de Processos com Agrupamento Inteligente

Os processos classificados como analisáveis são por sua vez ranqueados usando técnicas de agrupamento por aprendizagem de máquina. O algoritmo utilizado para tal análise foi o K-means [4].

O objetivo do ranqueamento é identificar o grau de relevância dos processos e separá-los em grupos de maior interesse para o estado. Uma vez treinado o algoritmo e encontrados os parâmetros que definem os grupos, esse algoritmo não precisa ser novamente treinado, a menos que haja alteração nos termos chave ou no fluxo de trabalho. No trabalho cotidiano, esse algoritmo já treinado, deve ser utilizado para agrupar novos processos.

III. RESULTADOS

A. Classificador Inteligente de Documentos

O classificador inteligente de documentos foi treinado com um banco de 2059 processos do ano de 2016. Sendo desses, 1062 rotulados como importantes e 997 como não importantes. Como documentos importantes estão textos jurídicos do tipo “petição inicial” e “primeiras declarações”. Como não importantes estão textos dos seguintes tipos: “despacho”, “protocolo”, “sentença”, “extrato processual”, “ofício”, “decisão”, “documentos de identificação pessoal”, “intimação” e “certidão de óbito”. Um quarto desses documentos foram retirados para validação; dos restantes, um quarto foi retirado para teste, sobrando aproximadamente 56% dos arquivos para treino. Como resultados, obteve-se uma acurácia de 98,4% para teste e 98,0% para treino.

B. Ranqueamento Inteligente de Processos

Para ranqueamento dos processos, foram avaliados 5171 processos datados entre 2016 e 2019. Contudo, apenas 4868 processos restaram após as etapas de pré-processamento, onde foram retirados os documentos com idioma inapropriado; classificação inteligente de documentos, onde foram selecionados os documentos que continham a maior relevância por processos; pontuação semântica, onde foi calculada a pontuação para cada processo; e classificação condicional, na qual foram separados os documentos analisáveis.

A distribuição dos processos no espaço amostral de suas pontuações é mostrada na Fig. 2, na qual pode se ver uma grande concentração de processos na base, com baixa pontuação e uma pequena quantidade de processos espalhados pela região superior. Esses últimos podem ser considerados como *outliers* e por sua vez podem necessitar de uma atenção especial por parte da procuradoria.

Após essas etapas, foi realizado, com base nas pontuações, o ranqueamento de processos com agrupamento inteligente para os classificados como Analisáveis. A fim de determinar o melhor número de grupos, foi usada a técnica do cotovelo para avaliar a distribuição entre 1 e 10 grupos. Na Fig. 3 é mostrado o gráfico representativo para as 10 simulações, no qual, no seu eixo vertical, está a soma da distância quadrada (SDQ) entre o centroide e cada membro do grupo, que é usada para medir a

eficiência dos centroides escolhidos. Nesse gráfico, pode-se ver uma acentuação da curva entre os números de grupos 2 até 6. Neste último, a curva começa a reduzir sua inclinação, o que significa que o ganho por número de grupos está sendo cada vez menor. Sendo assim, foi escolhido o número de 5 grupos para ranquear os processos.

Esse ranqueamento pode ser visualizado no gráfico de dispersão na Fig. 4, os grupos, por sua vez, foram numerados da esquerda para direita, indo do 1 ao 5. As pontuações mínimas e máximas encontradas pelo algoritmo inteligente para cada grupo, assim como a quantidade de processos por grupos, são mostrados na Tabela 1.

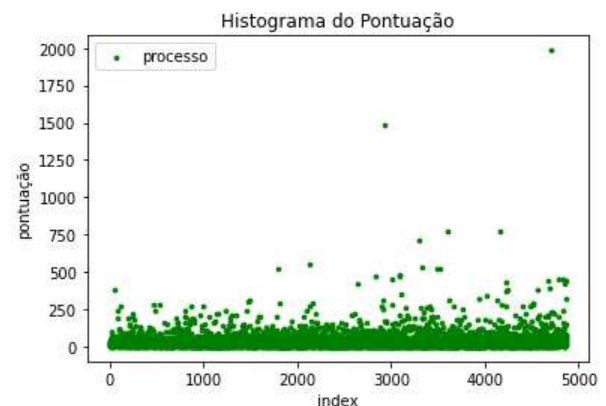


Figure 2. Histograma de Pontuação de Processos

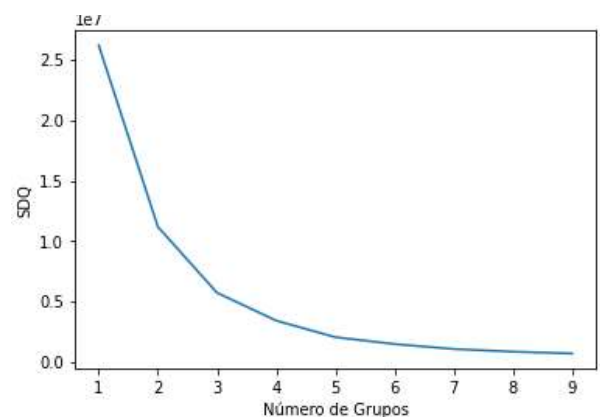


Figure 3. Gráfico de Cotovelo para o Agrupamento

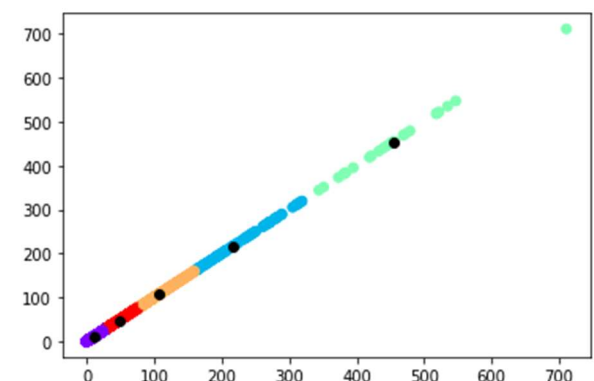


Figure 4. Gráfico de Agrupamento de Processos

TABLE I. RESULTADOS DO RANQUEAMENTO

Grupo	Pontuação Mínima	Pontuação Máxima	Quantidade de Processos
1	0	29	2939
2	30	76	1351
3	77	160	430
4	162	319	120
5	344	1985	29

IV. CONCLUSÕES

Dada a elevada quantidade de processos e consequente sobrecarga dos procuradores, o que acarreta em perdas financeiras diretas e indiretas para o Estado, a atual abordagem apresentou um sistema inteligente para suporte a decisão que automatiza todo o processo de coleta, transformação, análise e ranqueamento de processos de herança. Para isso foram utilizadas técnicas de OCR, processamento de imagem, mineração de texto, classificação inteligente de textos, pontuação semântica e agrupamento inteligente.

Após o ranqueamento dos processos disponíveis na base de dados, os intervalos de pontuação por grupo encontrados podem ser usados para agrupar novos processos sem a necessidade de um novo treinamento do algoritmo. Uma vez recebidos, rotineiramente, novos processos no setor de análise de heranças, esses podem ser automaticamente ranqueados, seguindo o fluxo estabelecido e distribuído aos colaboradores de acordo com a relevância dos processos e o grau de especialidade dos analistas. Dessa forma, processos mais importantes, que são os dos grupos 3, 4 e 5 podem ser distribuídos entre os procuradores, os quais são os funcionários de maior nível dentro da procuradoria e que estão em menor número. Os processos do nível 2 podem ser repassados aos analistas auxiliares de grau sênior ou pleno, enquanto que os processos de nível 1, os mais numerosos e menos relevantes, podem ser atribuídos aos analistas júnior ou em formação, resultando em uma força tarefa mais numerosa e de menor custo para o estado.

Além disso, durante as etapas de pré-processamento e classificação condicional, alguns processos são descartados devido à má qualidade dos arquivos de imagem, resultando em idiomas não identificados, processos duvidosos e incompletos. Esse tipo de processo necessita de uma operação técnica pouco especializada para que retornem ao sistema, atividade que pode ser distribuída entre técnicos administrativos, diminuindo a quantidade de trabalho destinada aos profissionais de maior especialidade. Os documentos classificados como duvidosos, por sua vez, podem indicar novos termos-chaves a serem acrescentados após análise dos especialistas.

O fluxo proposto também contribui com suporte à decisão na análise de processos, uma vez que indica os termos-chaves encontrados por processo e a pontuação total atribuída. Essas informações podem ser usadas para facilitar a interpretação dos textos jurídicos e acelerar o processo de análise. Essa abordagem permite que o especialista possa inserir sua experiência e ponto de vista na análise, dando autonomia e confiabilidade ao sistema. Além do mais, isso torna o sistema

flexível, podendo ser adaptado para diferentes procuradores e setores apenas com uma rápida mudança personalizada nos termos-chave e suas pontuações, quando necessário.

A sugestão de ranqueamento de processos, redistribuindo o trabalho de acordo com o grau de instrução dos colaboradores pode aumentar a arrecadação do estado, uma vez que os processos de maior importância monetária podem ser devidamente analisados por profissionais mais qualificados, que estarão mais disponíveis para oferecer a atenção requerida.

Com isso, é perceptível o grau de contribuição que o uso de técnicas de mineração de texto e de aprendizagem de máquina podem trazer para o setor jurídico público, aumentando a eficiência e rentabilidade da máquina pública, beneficiando a sociedade como um todo. Além do mais, a abordagem apresentada pode ser usada para automação de outros processos na administração pública.

Para consolidar o grau de melhoria da proposta, serão necessários estudos futuros a fim de calcular o retorno financeiro obtido em núcleos de herança que usam a abordagem proposta. Em trabalhos futuros, também pretende-se aumentar o grau de inteligência do processo com algoritmos de aprendizagem de máquina para análise semântica de palavras, o que pode trazer melhorias na escolha automática de termos-chaves e na atribuição adaptativa de pontuações.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001; e da Procuradoria Geral do Estado de Pernambuco.

REFERÊNCIAS BIBLIOGRÁFICA

- [1] Portal PGE, "Institucional", PGE. January 2021. Accessed on: February 16, 2021. [Online]. Available: <http://www.pge.pe.gov.br/pgeapresentacao.aspx>
- [2] Pernambuco, "Lei Nº 13.974", Legislação Tributária. Accessed on: February 16, 2021. [Online]. Available: https://www.sefaz.pe.gov.br/Legislacao/Tributaria/Documents/Legislacao/Leis_Tributarias/2009/Lei13974_2009.htm
- [3] A. L. G. Ferreira, "Princípio da Supremacia do Interesse Público e Princípio da indisponibilidade do interesse Público: Pilares do Regime Jurídico-Administrativo", in *Conteúdo Jurídico*, vol. 20, 2012.
- [4] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques", in *arXiv*, 2017, preprint arXiv:1707.02919.
- [5] Z. Ge, Z. Song, S. X. Ding and B. Huang, "Data mining and analytics in the process industry: The role of machine learning", in *IEEE Access*, vol. 5, 2017, pp. 20590–20616.
- [6] E. Souza et al. "Characterising text mining: a systematic mapping review of the Portuguese language", in *IET Software*, vol. 12, n. 2, 2017, pp. 49–75.
- [7] I. H. WITTEN et al., "Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2016.
- [8] J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)", in *IEEE Access*, vol. 8, 2020, pp. 142642–142668.
- [9] N. Gandhi and L. J. Armstrong, "A review of the application of data mining techniques for decision making in agriculture", in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, pp. 1–6, December 2016.

- [10] F. B. de Lima Neto, D. M. L. Martins and G. Vossen, "A semiotic-inspired machine for personalized multi-criteria intelligent decision support", in *Data & Knowledge Engineering*, vol. 117, 2018, pp. 225-238.
- [11] P. Chapman et al. "Crisp-dm 1.0 step-bystep data mining guide", in *SPSS inc*, vol. 9, 2000, pp. 13.
- [12] A. Mordvintsev and Abid K, "Image Thresholding", *OpenCV-Python Tutorials*. Accessed on: February 16, 2021. [Online]. Available: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html#otsus-binarization
- [13] XGBoost, "XGBoost Documentation". Accessed on: February 16, 2021. [Online]. Available: <https://xgboost.readthedocs.io/en/latest/>
- [14] Z. Qi, "The text classification of theft crime based on TF-IDF and XGBoost model". In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, pp. 1241-1246, June 2020.