Mineração de Texto para extrair informações de certidões federais e estaduais solicitadas em processos licitatórios

***Abstract*—Na era digital grande parte das informações encontram-se nos formatos não estruturados e semiestruturados, como os arquivos no formato PDF e páginas web. Pode-se citar, por exemplo, as certidões federais e estaduais que são solicitadas em processos licitatórios e disponibilizadas nos ambientes digitais, a maioria delas em formato PDF. Para habilitar as empresas para uma licitação é preciso analisar esses documentos de forma manual. Neste sentido, o objetivo deste estudo é apresentar a aplicação de técnicas de mineração de texto para extrair informações em certidões federais e estaduais do estado do Rio Grande do Norte – RN. No desenvolvimento deste trabalho foram realizadas pesquisa bibliográfica, estudo documental, análise das certidões, desenvolvimento do *script* de mineração de texto e exibição das informações. Como resultado possui um código capaz de minerar no total 12 certidões, 9 federais e 3 do estado do Rio Grande do Norte - RN, no entanto vale salientar que só foi possível extrair informações de 8 certidões federais e 2 do estado do RN. Assim sendo, percebe-se que a aplicação de Mineração de Texto nesses documentos, possibilita a estruturação, armazenamento e exibição dessas informações como data de validade, situação e**

***Index Terms*—licitação, mineração, certidões, situação.**

I. Introdução

Na era digital a quantidade de informações disponibilizadas é cada vez maior e a tendência é que esse processo torne-se cada vez mais acelerado. Convém salientar que, uma considerável parte dessas informações está no formato de texto, ou seja, na linguagem natural humana. De acordo com [1], essas informações podem ser encontradas em formatos não estruturados, (em que as sentenças são livres e/ou escritas em linguagem natural) ou semi estruturados (quanto não dispões de formatação rígida, permite variação na ordem dos dados e não respeitam rigidamente a gramática da linguagem natural, um exemplo é a abreviação de palavras), em modelos de documentos textuais, a exemplo dos arquivos no formato PDF e/ou páginas *web*.

Nesse contexto, grande parte das certidões federais e estaduais são disponibilizadas nos ambientes digitais. Para habilitar uma empresa em processos licitatórios é necessário reunir e analisar, de forma manual (baixar e ler respectivamente), inúmeras certidões que comprovem sua situação econômica, financeira ou fiscal. Neste sentido, a atividade de recuperação de informações manualmente a partir de textos pode ser uma tarefa maçante e repetitiva, que demanda grande quantidade de tempo para quem a realiza, estando sujeito ainda à ocorrência de falha humana na identificação.

Assim sendo, para facilitar e auxiliar a atividade de procura e seleção das informações relevantes, pode-se utilizar técnicas de mineração de textos que, segundo [1], é a área que usa algoritmos para processar textos e identificar informações úteis, que geralmente não poderiam ser recuperadas por métodos tradicionais de consultas. Segundo [3], "a mineração de textos consiste na descoberta da informação através da extração de dados a partir de coleções de textos dos mais variados tipos". Mediante ao exposto, o objetivo deste trabalho é apresentar a aplicação de técnicas de mineração de texto para extrair informações em certidões federais e estaduais do estado do Rio Grande do Norte - RN. Estas certidões são utilizadas por empresas para habilitação em processos licitatórios. Esse texto está organizado a partir desta introdução, seguido da seção 2, com o referencial teórico; na seção 3 são apresentados os materiais e métodos; posteriormente, na seção 4, são abordados os resultados; e, por fim, na seção 5 são apresentadas as considerações finais.

II. Referencial Teórico

Esta seção tem como objetivo apresentar conceitos importantes que fundamentam o desenvolvimento do presente estudo, como a noção de mineração de texto. Para mais, tenciona apresentar trabalhos relacionados a essa temática.

A. Mineração de Texto

A Mineração de Texto (doravante MT), para [2], é considerada como uma subárea da mineração de dados. A principal diferença entre ambas é que a primeira trabalha com dados em formato de linguagem natural não estruturados, enquanto a segunda, por sua vez, é aplicada a dados estruturados. De acordo com [4],os dados em formato de texto não possuem uma estrutura, pois são criados diretamente pelos usuários, isso culmina na dificuldade de tratamento e análise em grande quantidade, um exemplo desses dados são as informações das redes sociais. Já [5] afirma que a MT torna mais fácil a visualização de informações que são difíceis de serem percebidas.

Na concepção de [6], a MT se relaciona com diversas áreas, como: aprendizado de máquina, recuperação de informações, processamento de linguagem natural e estatística. Além disso, conforme [7], pode ser utilizada comercialmente, tendo em vista que pode ser aplicada em diferentes áreas do conhecimento.

Conforme [1], as contribuições da MT se referem à busca por dados úteis e ao entendimento de conteúdos em documentos de distintos formatos, a exemplo de *e-mails*, páginas *web*

e arquivos no formato PDF. À vista disso, para [8], a MT não produz conhecimento, mas facilita a construção dele, por meio da busca e análise de informações contidas em um ou mais documentos.

De acordo com [2], há dois tipos de MT: análise estatística e semântica. A primeira tem como foco a frequência de termos, enquanto a segunda faz um exame da sequência de termos no contexto de uma frase. Para realizar essa análise é necessário o uso de técnicas de Processamento de Linguagem Natural.

Um método utilizado para extrair essas informações é o uso de expressões regulares, que, segundo [9], "são uma forma de notação para descrever a língua (conjunto de *strings* produzidas), sendo usadas por diversos editores de texto e linguagens de programação para procurar linguagens de *scripting*, manipulando o texto com base em padrões".

Ademais, o processo de MT, conforme [10], pode ser dividido em coleta, pré-processamento, indexação, mineração e análise. Nos próximos tópicos, essas fases serão abordadas.

1) Coleta: De acordo com [11], a coleta é imprescindível para a extração dos textos que servirão para as próximas etapas do processo de MT. Em relação aos dados, eles podem estar em distintos locais, como banco de dados, dispositivos de armazenamento de memória (HD) e na própria *internet*.

Na *internet*, existe uma grande quantidade de informações que podem ser coletadas de forma automática. Na concepção de [11], uma das formas é a *Web Crawlers*, constituída por robôs capazes de buscar esses dados, podendo identificar em uma página HTML apenas seu conteúdo de texto. Além dessa forma, ainda existe*Web Scraping*, que, nas palavras de [12], refere-se à extração de informações em *sites* por meio da estrutura sintática de códigos HTML para detectar, selecionar e coletar os dados úteis.

Neste estudo utilizou-se a *Web Scraping* para acessar os *sites* e fazer o *download* das certificações. Posteriormente, foi feita a conversão do arquivo PDF em formato de texto para poder ser realizado o pré-processamento desses dados.

2) Pré-processamento: Conforme [13], o pré-processamento é a atividade responsável por tratar a grande quantidade de textos que foram obtidos durante a fase de coleta, para que eles possam se tornar adequados e gerar melhores resultados.

Segundo [11], quando o texto é extraído, ele pode vir de forma não estruturada; assim sendo, existe a necessidade de que ele seja tratado e que não perca o sentido. Para o autor, essa etapa é denominada pré-processamento e tem como objetivo realizar a preparação dos dados que serão processados na fase de mineração.

Para mais, um fator importante durante o pré-processamento é a criação dos *tokens*, que [14] afirma ser "um primeiro passo para o processamento de um texto, sendo uma etapa crucial de segmentação da informação". Na realidade, trata-se de um fracionamento de palavras, delimitadas por *caracteres* pré-determinados. Cada unidade separada é denominada *token*.

Em conformidade com [15], a tokenização tem como propósito desmembrar um documento textual em unidades mínimas, de modo a manter o significado do texto original. Ainda de acordo com [15], a tokenização é utilizada quando está sendo realizado o processamento de linguagem natural, como forma de segmentação das palavras para quebrar a sequência de *caracteres* em um texto, localizando o limite de cada palavra ou sentenças de unidade.

3) Mineração: Quanto à mineração, [11] afirma que essa etapa compreende a definição e a aplicação de um ou mais algoritmos de extração de conhecimento. Ademais, convém lembrar que, quando se chega nessa fase, os dados já apresentam uma melhor estruturação, para que assim seja realizado o processamento principal. No que tange à extração das informações desejadas, o autor reitera que o algoritmo de mineração é aplicado no conjunto de tokens que foram criados na etapa anterior.

4) Interpretação ou Pós-processamento: Para [10], a fase de interpretação ou pós-processamento é considerada como a interpretação dos resultados obtidos nas etapas anteriores, em que o analista dos dados verifica se o classificador atingiu os objetivos esperados, por meio de métricas como a taxa de erros e a complexidade do modelo.

Na perspectiva de [11], é nessa fase que são exibidos os resultados de todo o processo, desde a fase de coleta, pré-processamento e mineração, de forma clara e eficiente, para que possam ser extraídas as informações desejadas. Porém, se elas não forem suficientes, é possível voltar para as fases antecedentes. Além disso, para a exibição dos resultados obtidos na MT, pode-se usar, por exemplo, gráficos, tabelas e aplicações *web*.

B. Trabalhos Relacionados

Nesta seção, são apresentados trabalhos relacionados ao assunto MT, que é o principal propósito deste estudo.

Em sua pesquisa, [16] usou a MT no âmbito da saúde, em especial, nos prontuários eletrônicos de pacientes, para extrair informações a respeito de anamnese, que são divulgadas no ciberespaço. Ressalta-se que a aplicação dessa técnica encontrou muitas dificuldades, visto que a anamnese tem distintas formas de ser redigida.

No seu trabalho, [17] utilizou a MT para classificar processos judiciais voltados para a área trabalhista em 241 documentos do tipo Recursos Ordinários, extraídos de processos do PJe instalado na Justiça do Trabalho. Para isso, usou o comparativo de distintos algoritmos, obtendo, com o algoritmo *Multi-Layer Percepton*, 46,03% de precisão em relação ao assunto principal. Mesmo com a precisão de identificação do assunto um pouco baixa, comprovou-se ser possível extrair conhecimento desse tipo de documento.

Já [12], por sua vez, usou técnicas de MT em dados relacionados a *fake news* na área da saúde, mais precisamente em textos do Portal da Saúde. Como se sabe, notícias falsas podem gerar insegurança para a sociedade, e esse fato mostra a relevância de estudá-las. Acresce a isso que investigá-las através dessas técnicas evidencia como a MT é útil aos diversos campos de pesquisa, visto que *fake news* é um tema que afeta diferentes ciências.

A pesquisa de [18] fez um comparativo com algoritmos de comparação de métodos de MT aplicado à classificação de documentos jurídicos. Para tanto, foram usados algoritmos de aprendizado de máquina supervisionado de jurisprudência. Assim sendo, com essas técnicas de classificadores e aprendizado de máquina, é possível extrair e exibir as informações de documentos.

Por fim, o estudo de [19] buscou, por meio da MT, a obtenção de conhecimento em documentos/arquivos atinentes à investigação policial com o uso de diferentes técnicas. Nesse sentido, apresentou bons resultados quanto ao achado de conhecimentos sobre entidades, conexões e categorias temáticas. Por conseguinte, percebe-se que as técnicas de MT podem ser utilizadas para recuperação de informações em distintas áreas, como: jurídicas, policiais e na saúde.

III. Materiais e Métodos

O presente estudo foi desenvolvido em três etapas: na primeira, realizou-se uma revisão bibliográfica; na segunda, um estudo documental; e, na terceira, desenvolveu-se a aplicação responsável pela extração das informações.

A pesquisa bibliográfica foi efetuada por meio de textos que já foram divulgados, o que corrobora a conceituação exposta por [20], para quem esse tipo de pesquisa é fundamentado em materiais já publicados. Esses materiais podem ser livros, revistas, jornais, teses, dissertações e anais de congressos científicos. Dessa forma, realizou-se uma pesquisa bibliográfica de trabalhos relativos aos seguintes temas: MT e trabalhos relacionados.

Sobre o estudo documental, [21] afirma que ele possui como fonte de coleta de dados/informações apenas documentos, escritos ou não, que são considerados como fontes primárias. As fontes documentais podem ser arquivos públicos, particulares e estatísticos.

Assim, com o fito de identificar quais certidões são requisitadas em processos licitatórios, foi realizada a busca e análise de cerca de 10 editais, dos quais se podem citar: Edital Pregão Eletrônico nº 6/2021-0002, da Prefeitura Municipal de Pau dos Ferros/RN; Edital Pregão Eletrônico nº 14/2021, do Tribunal de Contas do Estado do Rio Grande do Norte (TCE/RN); Edital Pregão eletrônico nº 18/2021, do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, Polo Reitoria . Na Fig. 1, são listadas doze certidões que podem ser encontradas em *sites* governamentais na esfera federal e do estado do RN e que podem ser realizadas a mineração de texto.

Quanto ao desenvolvimento do *script* de extração, recorreu-se à linguagem de programação *python*, à biblioteca Pymupdf – para conversão dos arquivos PDF em *strings* (texto) –, à *Regex* (ou Re) – para realizar as consultas no texto convertido – Pandas na para manipulação dos dados extraídos – e, para auxiliar, ao *IDE VSCode* e ao *Colab*.

IV. Resultados

Nesta seção, serão apresentados os resultados obtidos durante a aplicação da metodologia/técnica de MT para a coleta

Fig. 1. Lista de Certidões.

de um conjunto de dados nas certidões federais do estado do Rio Grande do Norte (RN). A Fig. 2 apresenta o diagrama de atividades, com todas as ações necessárias para o processo de extração de informações das certidões.

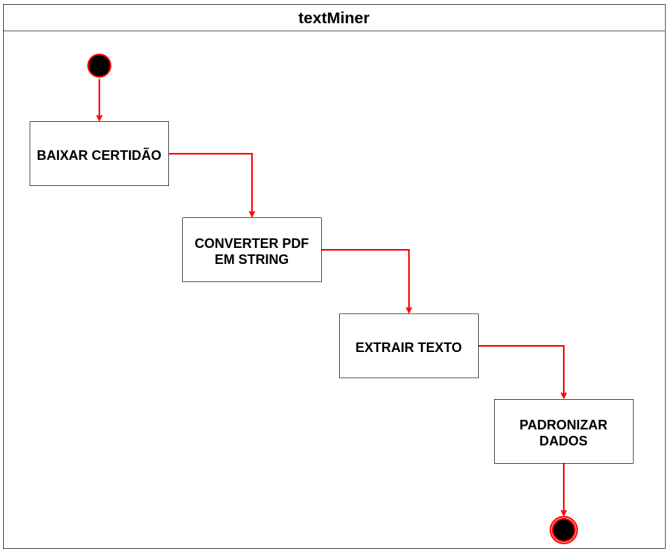

Fig. 2. Diagrama de atividades.

O diagrama de atividades é formado por 4 ações. A primeira delas refere-se ao *download* manual do documento, que foi utilizado posteriormente para ter acesso à certidão, foi preciso informar o CNPJ. A segunda atividade, por sua vez, consiste em converter o PDF em *String*, utilizando a biblioteca *PyMuPDF*, que realiza essa conversão e retorna o texto contido no arquivo PDF. A terceira ação relaciona-se à extração do texto. Esse método tem como propósito utilizar a MT, em especial expressões regulares, para buscar informações na *String* que foi criada na etapa anterior. Por fim, destaca-se a etapa de padronização dos dados. Nesse caso, utilizam-se funções para realizar a padronização das informações, como datas do padrão 00/00/1999 para um 2022-04-28, por exemplo, para que posteriormente possam ser armazenadas em uma base de dados. A Fig. 3 exibe as informações referentes às dez certidões adquiridas e que passaram pelo processo de MT. Esses dados dizem respeito ao nome da certidão (essa aqui não foi obtida por meio de MT), data de validade e à situação.

	Nome	Data de Validade	Situação
0	Cadastro Nacional de Pessoa Jurídica	2022-04-28 00:00:00+00:00	Regular
1	Certidao Conjunta Negativa de Débitos Relativo...	2022-07-27 00:00:00	Regular
2	Consulta Consolidada de Pessoa Jurídica	2022-04-28 00:00:00+00:00	Regular
3	Certidão de Débito Trabalhista - CDT	2022-09-26 00:00:00	Regular
4	Certidão de Inscrição do Estado do Rio Grande ...	2022-04-28 00:00:00+00:00	Irregular
5	Cadastro Nacional de Condenações Cíveis por At...	2022-05-02 00:00:00+00:00	Regular
6	Certidão Negativa de Débitos Relativos Aas Tri...	2022-06-07 00:00:00	Regular
7	Certidão Negativa de Licitantes Inidôneos	2022-04-28 00:00:00+00:00	Regular
8	Certidão de Débito Tabalhista	2022-04-28 00:00:00+00:00	Irregular
9	Certificado de Regularidade do FGTS	2022-04-24 00:00:00	Regular

Fig. 3. Informações obtidas das certidões.

Salienta-se, contudo, que as certidões de Microempreendedor Individual (MEI) e Certidão Estadual de Falência e/ou Recuperação Judicial não foram adquiridas – no caso das primeiras, porque o CNPJ utilizado pertence ao IFRN, que é uma entidade pública. Porém, é possível baixar a certidão informando um CNPJ de uma empresa. Sobre a Certidão de Falência, o sistema requer que o usuário a solicite para que ela chegue ao e-mail e, ao solicitar, não chegou nenhuma mensagem informando se a solicitação foi aprovada ou não. Assim, não foi possível realizar a MT nesses casos.

V. Considerações Finais

Este trabalho apresentou a aplicação de MT em certidões federais e do estado do RN, tencionando realizar a recuperação de informações. Ademais, salienta-se que essas certidões são geralmente solicitadas em processos licitatórios das esferas federal, estadual e municipal, e que as empresas devem analisar essa documentação de forma manual, para verificar se esses documentos estão regulares.

Diante do exposto, percebe-se, com a aplicação dessa técnica nesses documentos, a possibilidade de estruturação das informações obtidas, armazenamento e exibição desses dados em aplicações *web* e/ou móvel, para que os usuários possam visualizar, de forma mais precisa, apenas as informações que lhes convêm, facilitando todo o processo de análise das certidões. Quanto a trabalhos futuros, pretende-se ampliar a MT para um número maior de certidões, em especial, abrangendo toda a região Nordeste e, posteriormente, as demais regiões brasileiras.

References

[1] ZABOT, Guilherme Felipe. Uma aplicação de Mineração de Texto em Faturas de Telefonia Móvel Corporativa. Cascável/PR: 2016. 82 p. Trabalho de Conclusão de Curso (Graduação). Universidade Estadual do Oeste do Paraná, Curso de Bacharelado em Ciência da Computação, Centro de Ciências Exatas e Tecnológicas. Cascavel/PR, 2016.

[2] TORRES CARREÑO, L. A. et al. Criação de modelos de preferência a partir de textos da web: caso de aplicação no segmento de vestuário= Creation of preference models from web texts: a case study in fashion sector. Limeira/SP, 2018. 71 p. Dissertação (Mestrado). Universidade Estadual de Campinas, Faculdade de Tecnologia, Limeira/SP.

[3] SOARES, V. D. S. MINERAÇÃO DE TEXTOS PARA IDENTIFICAR PERFIS DE SATISFAÇÃO DE CLIENTES. 2016. 56 f. TCC (Graduação) - Curso de Curso de Ciência da Computação, Universidade de Santa Cruz do Sul, Santa Cruz do Sul, Rs, 2016.

[4] MARCOLIN, Carla et al. (2020). "ARGUMENTOS DA DECISÃO DE VOTO DE DEPUTADOS DURANTE A VOTAÇÃO DO IMPEACHMENT". In: IEEE International Conference on Big Data (Big Data). 2020.

[5] OLIVEIRA, R. J. V.; Silva, R. S. D.; Gava, T. B. S. (2018). USO DE SOFTWARE LIVRE PARA A MINERAÇÃO DE TEXTO. 2018.

[6] CAVALCANTI, R. D. (2017). Classificação de tendências políticas em notícias via mineração de texto e redes neurais sem peso. Rio de Janeiro, 2017.

[7] PEZZINI, A. (2017). Mineração de textos: conceito, processo e aplicações. Revista Eletrônica do Alto Vale do Itajaí, v. 5, n. 8, p. 58-61, 2017.

[8] OLIVEIRA, S. D. (2017). O processo de construção da coerência textual na escrita acadêmica com base na mineração de texto. 2017.

[9] MOTA, F. A. T. (2020). Recolha de Informação de Eventos Desportivos usando Técnicas de Mineração de Dados: investigação, desenho e implementação de uma solução web baseada na nuvem. 2020. 91 f. Tese (Doutorado) - Curso de Engenharia Informática, Informatica, Universidade Beira Interior, Covilhã, 2020.

[10] ARANHA, C. N. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. Rio de Janeiro/RJ, 2007. 144 p. Tese (Doutorado). Pontifícia Universidade Católica do Rio de Janeiro, Curso de Engenharia Elétrica, Departamento de Engenharia Elétrica, Rio de Janeiro, 2007.

[11] SANTOS, W. P. S. Análise dos tweets sobre a black friday através da mineração de texto e análise de sentimentos. Rio de Janeiro/RJ, 2016. 51 p. Trabalho de Conclusão de Curso (Graduação). Universidade Federal do Estado do Rio de Janeiro, Centro de Ciências Exatas e Tecnologia Escola de Informática Aplicada. Rio de Janeiro, 2016.

[12] VIEIRA, L. M.; SILVA, N. R.; CORDEIRO, D. F. Análise descritiva das fake news da saúde através de mineração de textos no Portal da Saúde. Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, XXI Congresso de Ciências da Comunicação na Região Centro-Oeste, Goiânia/GO, 2019. Disponível em: https://portalintercom.org.br/anais/centrooeste2019/resumos/R66-0230-1.pdf. Acesso em: 03 set. 2021

[13] CAPOBIANCO, K. R. Avaliação da etapa de pré-processamento na mineração de texto em redes sociais digitais. Londrina/PR, 2016. 57 p. Trabalho de Conclusão de Curso (Graduação). Universidade Estadual de Londrina, Curso de Ciência da Computação. Londrina/PR, 2016.

[14] SAKURAI, G. Y. Processamento de linguagem natural - detecção de fake news. Londrina/PR, 2019. 37 p. Trabalho de Conclusão de Curso (Graduação). Universidade Estadual de Londrina, Curso de Ciência da Computação. Londrina/PR, 2019.

[15] RODRÍGUEZ, M. M.; BEZERRA, B. L. D. Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (Portarias). Revista de Engenharia e Pesquisa Aplicada, Edição Especial, p. 67-77, 2020. Disponível em: http://revistas.poli.br/index.php/repa/article/view/1204/576. Acesso em: 03 set. 2021.

[16] CARVALHO, R. C. D. (2017). Aplicação de técnicas de mineração de texto na recuperação de informação clínica em prontuário eletrônico do paciente. 2017.

[17] ROCHA, A. C. P. (2019). Mineração de textos para classificação de processos judiciais trabalhistas. 2019.

[18] SILVA, E. C. M. D.; Medeiros, B. A. (2020). Comparação de métodos de mineração de texto para classificação de documentos jurídicos. Ciência da Computação-Tubarão, 2020.

[19] SILVA, M. P. D.; Viera, A. F. G. Descoberta de conhecimento com uso de técnicas de mineração de textos aplicadas em documentos textuais da investigação policial brasileira. Investigación bibliotecológica, v. 35, n. 88, p. 161-183, 2021.

[20] GIL, A. C. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, 2002.

[21] LAKATOS, E. M. (2017). Fundamentos de metodologia científica / Marina de Andrade Marconi, Eva Maria Lakatos. – 8. ed. – São Paulo: Atlas, 2017.