DOI: http://dx.doi.org/10.20435/multi.v30i74.4727
Recebido em: 23/10/2024; aprovado para publicação em: 06/11/2024

Mineração de texto aplicada à análise de estudos de impacto ambiental de pequenas centrais hidrelétricas do Estado de Mato Grosso

Text mining applied to the analysis of environmental impact studies of small hydroelectric plants in the State of Mato Grosso

Minería de texto aplicada al análisis de estudios de impacto ambiental de pequeñas centrales hidroeléctricas en el Estado de Mato Grosso

Lucas Michelotti Baldini[1]
Anderson Castro Soares de Oliveira[2]
Ibraim Fantin da Cruz[3]
Lia Hanna Martins Morita[4]

[1] Mestre em recursos hídricos pela Universidade Federal de Mato Grosso (UFMT) e coordenador geral da Faculdade Católica de Cuiabá. **E-mail:** michelotti.lucas@gmail.com, **ORCID:** https://orcid.org/0009-0005-7628-0349

[2] Doutor em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras. Professor do departamento de estatística da Universidade Federal de Mato Grosso (UFMT). **E-mail:** anderson.oliveira@ufmt.br, **ORCID:** https://orcid.org/0000-0001-6222-9300

[3] Doutor em Recursos Hídricos e Saneamento Ambiental pelo Instituto de Pesquisas Hidráulicas da Universidade Federal do Rio Grande do Sul (UFRGS) e professor do departamento de engenharia sanitária e ambiental da Universidade Federal de Mato Grosso (UFMT). **E-mail:** ibraim.cruz@ufmt.br, **ORCID:** https://orcid.org/0000-0001-6731-0036

[4] Doutora em Estatística pela Universidade Federal de São Carlos (UFSCar) e professora do departamento de estatística da Universidade Federal de Mato Grosso (UFMT). **E-mail:** lia.morita@ufmt.br, **ORCID:** https://orcid.org/0000-0001-5912-5754

Resumo: Com o avanço das legislações ambientais e das técnicas de controle de impactos, a quantidade de dados gerados aumentou significativamente, mas muitos desses dados são subutilizados. A mineração de texto, que extrai informações valiosas de dados textuais não estruturados, surge como uma ferramenta essencial para analisar relatórios de impacto ambiental e identificar padrões que podem melhorar políticas públicas e estratégias de gestão. Este estudo focou na análise textual dos estudos de impacto ambiental (EIA) de pequenas centrais hidrelétricas (PCHs) em Mato Grosso, utilizando técnicas como mineração de texto, nuvens de palavras e grafos para identificar temas recorrentes e inter-relações nos textos. A pesquisa também ressaltou a importância do pré-processamento dos textos para garantir análises precisas e a necessidade de uma revisão das legislações e melhorias na organização dos documentos disponíveis no site da SEMA-MT, para facilitar o acesso e a consulta pública.

Palavras-chave: Processamento de Linguagem Natural (PLN); pré-processamento; frequência de palavras.

Abstract: With the advancement of environmental regulations and techniques to control environmental impacts, the amount of data generated has significantly increased, but much of this data is underutilized. Text mining, which extracts valuable information from unstructured textual data, emerges as an essential tool for analyzing environmental impact reports and identifying patterns that can improve public policies and management strategies. This study focused on the textual analysis of Environmental Impact Studies (EIA) for Small Hydroelectric Plants (PCHs) in Mato Grosso, using techniques such as text mining, word clouds, and graphs to identify recurring themes and interrelationships in the texts. The research also highlighted the importance of text preprocessing to ensure accurate analysis and the need for a revision of regulations and improvements in the organization of documents available on the SEMA-MT website to facilitate access and public consultation.

Keywords: Natural Language Processing (NLP); preprocessing; word frequency.

Resumen: Con el avance de las legislaciones ambientales y de las técnicas de control de impactos, la cantidad de datos generados ha aumentado significativamente, pero muchos de estos datos son subutilizados. La minería de texto, que extrae información valiosa de datos textuales no estructurados, surge como una herramienta esencial para analizar informes de impacto ambiental e identificar patrones que pueden mejorar las políticas públicas y estrategias de gestión. Este estudio se centró en el análisis textual de los Estudios de Impacto Ambiental (EIA) de Pequeñas Centrales Hidroeléctricas (PCH) en Mato Grosso, utilizando técnicas como minería de texto, nubes de palabras y grafos para identificar temas recurrentes e interrelaciones en los textos. La investigación también destacó la importancia del preprocesamiento de los textos para garantizar análisis precisos y la necesidad de una revisión de las legislaciones y mejoras en la organización de los documentos disponibles en el sitio de SEMA-MT, para facilitar el acceso y la consulta pública.

Palavras clave: Procesamiento de Lenguaje Natural (PLN); preprocesamiento; frecuencia de palavras.

1 INTRODUÇÃO

Com o avanço das legislações ambientais e o desenvolvimento de novas técnicas de controle dos impactos ambientais, a quantidade de informações geradas tem crescido de forma exponencial. Grande parte desses dados é utilizada exclusivamente para fins legais e oficiais, o que frequentemente resulta na subutilização de um vasto potencial de conhecimento contido nesses documentos. Essas informações, muitas vezes arquivadas sem serem totalmente exploradas, contêm *insights* valiosos que poderiam ser aproveitados para aprimorar políticas públicas, melhorar a tomada de decisões e criar novas estratégias de gestão ambiental. No entanto, com os avanços científicos e tecnológicos, surge a oportunidade de resgatar esse conhecimento oculto, transformando dados brutos em informações organizadas e úteis.

A mineração de texto, uma técnica que permite a extração de conhecimento a partir de dados textuais não estruturados, tem se mostrado essencial na era digital, dada a enorme quantidade de informações disponíveis em formato textual. Esta técnica, que utiliza algoritmos de processamento de linguagem natural (NLP) e métodos estatísticos, possibilita a identificação de padrões, tópicos e relações em textos, o que facilita a compreensão e a análise de grandes volumes de dados. Aplicações práticas da mineração de texto são vastas e incluem desde a análise de satisfação de clientes, categorização de conteúdo e detecção de fraudes, até a extração de informações em contextos jurídicos.

No contexto da gestão ambiental, a mineração de texto desempenha um papel crucial ao permitir a análise detalhada de dados ambientais e hidrológicos. Esta técnica pode ser utilizada para examinar relatórios de impacto ambiental, identificar tendências em estudos de caso e comparar a eficácia de diferentes abordagens de mitigação de impactos ambientais. Documentos como os estudos de impacto ambiental (EIAs), que frequentemente são exigidos para o licenciamento de projetos como pequenas centrais hidrelétricas (PCHs), contêm um volume significativo de informações textuais que, quando analisadas de maneira sistemática, podem revelar padrões e discrepâncias que não são facilmente detectáveis por meio de métodos tradicionais. Assim, a mineração de texto se torna particularmente

relevante para o gerenciamento e a análise de grandes volumes de documentos oficiais, oferecendo uma compreensão mais profunda dos impactos ambientais e contribuindo para a formulação de políticas mais eficazes.

Atualmente, a mineração de texto é uma ferramenta multidisciplinar que encontra aplicação em diversas áreas do conhecimento, incluindo a hidrologia e a gestão de recursos ambientais. Combinando tecnologias de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados, a mineração de texto oferece uma nova perspectiva sobre a análise de documentos complexos como os EIAs. Essa abordagem não apenas contribui para a melhoria dos processos de licenciamento e monitoramento ambiental, mas também ajuda a identificar áreas que necessitam de maior atenção, tornando-se uma ferramenta indispensável na gestão ambiental moderna.

Considerando o contexto específico das PCHs em Mato Grosso, este trabalho tem como objetivo classificar os EIAs dessas PCHs por meio da análise textual. Ao focar na organização textual dos EIAs disponíveis de forma digital no *site* da Secretaria Estadual de Meio Ambiente de Mato Grosso (SEMA-MT), busca-se explorar as similaridades e discrepâncias nos diagnósticos de impactos ambientais, com um enfoque particular na mineração de texto do capítulo de impactos ambientais dos documentos estudados. Esta análise não apenas melhorará a compreensão dos impactos específicos das PCHs no estado, mas também contribuirá para a melhoria das práticas de licenciamento e gestão ambiental, promovendo a sustentabilidade e a mitigação dos impactos ambientais desses empreendimentos.

2 REFERENCIAL

2.1 Estudo de Impacto Ambiental e Relatório de Impacto Ambiental

A Lei Federal n. 6.938, de 1º de agosto de 1981 (Brasil, 1990), estabelece a Política Nacional do Meio Ambiente no Brasil e marca o início das discussões sobre os impactos ambientais do desenvolvimento. O EIA

e o Relatório de Impacto Ambiental (RIMA) foram formalmente definidos pelo Decreto Federal n. 99.274, de 6 de junho de 1990 (Brasil, 1990), e são regulamentados pelas Resoluções do CONAMA n. 01, de 23 de janeiro de 1986, e n. 237, de 19 de dezembro de 1997 (Brasil, 1986; 1997). Estes documentos técnicos avaliam os impactos potenciais de projetos no meio ambiente, com o EIA detalhando os impactos e o RIMA apresentando uma versão simplificada ao público e aos órgãos públicos.

O EIA-RIMA é fundamental para o licenciamento ambiental, fornecendo uma análise detalhada dos efeitos ambientais e permitindo que o órgão responsável tome decisões informadas sobre a concessão de licenças. A clareza e a acessibilidade dessas informações são cruciais para garantir a participação pública e a transparência no processo.

A estrutura e a implementação do EIA-RIMA enfrentam desafios e exigem procedimentos rigorosos, conforme destacado por estudiosos como Agra Filho e Moura. A eficácia da avaliação de impactos e a governança ambiental são essenciais para promover a sustentabilidade e a responsabilidade socioambiental, refletindo a necessidade de constante atualização das metodologias e regulamentações para enfrentar os desafios ambientais atuais.

2.2 Mineração de Texto

A mineração de texto é parte integrante de um conceito mais amplo, conhecido como Knowledge Discovery in Text (KDT), que, em tradução livre, significa "descoberta de conhecimento em textos". O KDT refere-se ao processo de extrair informações valiosas a partir de grandes volumes de dados textuais não estruturados, utilizando técnicas específicas de mineração de texto. Esse processo visa identificar padrões ocultos, temas e *insights* que podem subsidiar a tomada de decisões em diversos contextos, como empresarial, científico e público (Srivastava; Sahami 2009; Ur-Rahman, 2017).

O KDT é uma adaptação do Knowledge-Discovery in Databases (KDD), descoberta de conhecimento em bases de dados, que tradicionalmente lida com dados estruturados. No entanto, ao contrário do KDD, que opera sobre dados organizados em bases de dados, a mineração de texto foca em dados não estruturados, como documentos, artigos, relatórios, e-mails e outros

tipos de textos livres. Isso exige uma abordagem diferente, que combina métodos de processamento de linguagem natural (PLN), técnicas estatísticas e algoritmos de aprendizado de máquina para analisar e interpretar os textos (Srivastava; Sahami 2009; Ur-Rahman, 2017).

O processo de KDT geralmente segue etapas semelhantes às do KDD, incluindo:

1. Coleta de dados: identificação e extração de dados textuais relevantes.
2. Pré-processamento: limpeza dos dados, remoção de ruídos, palavras irrelevantes e aplicação de técnicas como a lematização (redução das palavras à sua forma básica).
3. Estruturação da informação: transformação dos textos em representações estruturadas, como matrizes termo-documento, para facilitar a análise.
4. Mineração de texto: aplicação de algoritmos de mineração que buscam padrões, temas recorrentes e associações.
5. Geração de conhecimento: extração de sequências de conhecimento que podem ser usadas para apoiar a tomada de decisões.

2.3 Processamento de Linguagem Natural (NLP)

O Processamento de Linguagem Natural (NLP, do inglês *Natural Language Processing*) é um campo de estudo que combina computação e técnicas estatísticas para extrair e compreender a linguagem natural dos seres humanos. É uma área de pesquisa essencial para a mineração de textos, pois permite manipular e interpretar textos escritos em linguagem natural, viabilizando a realização de tarefas como análise de sentimentos, tradução automática e reconhecimento de fala (Chowdhury, 2003).

O NLP tem se tornado uma ferramenta crucial para empresas que buscam melhorar a interação com seus clientes por meio de assistentes virtuais e *chatbots*, simulando uma comunicação quase humana. A implementação de NLP nesses sistemas possibilita a realização de tarefas complexas, como a resolução de problemas específicos de clientes por meio de *chat*, WhatsApp, ou *websites* corporativos (Turban *et al*., 2010).

O NLP pode ser dividido em quatro etapas principais:

6. Análise morfológica: define artigos, substantivos, verbos e adjetivos, armazenando-os em um dicionário.
7. Análise sintática: busca relacionamentos entre as palavras e verifica a estrutura gramatical, identificando sujeito, predicado, complementos etc.
8. Análise semântica: determina o sentido real das frases e palavras, eliminando ambiguidades.
9. Análise pragmática: integra as etapas anteriores para interpretar o significado completo das frases (Bulegon; Moro, 2010).

2.3.1 Pré-processamento de texto

O pré-processamento de texto é uma etapa fundamental na mineração de texto, pois as bases de dados geralmente contêm informações incompletas, ruidosas, dinâmicas e redundantes (Navega, 2002). Esta etapa inclui a limpeza dos dados, remoção de erros e preparação do texto para análise, como a remoção de *stopwords* (palavras que não contribuem significativamente para o sentido do texto, como artigos e preposições) e a padronização do texto para minúsculas.

A lematização, um método comum no pré-processamento de textos, reduz palavras às suas formas base, ou lemas, facilitando a normalização de variações morfológicas. Essa técnica melhora a precisão de modelos de Processamento de Linguagem Natural (PLN), ao unificar diferentes formas de uma palavra, reduzindo redundâncias e aumentando a consistência nas anotações sintática (Zizka; Darena; Svoboda, 2019).

O pré-processamento também envolve a construção de uma matriz documento-termo, essencial para a representação e análise dos textos. Técnicas de visualização como a nuvem de palavras são úteis para destacar as palavras mais frequentes e facilitar a compreensão dos dados textuais (Mohammad; Turney, 2013).

Nesse contexto, a mineração de texto apresenta-se como uma abordagem eficaz para identificar padrões e tópicos de interesse em grandes conjuntos de dados textuais, contribuindo significativamente para a extração

de conhecimento e a tomada de decisão informada em diversas áreas (Aykurt; Sesen, 2017; Feldman; Sanger, 2007; Zizka; Darena; Svoboda, 2019).

3 METODOLOGIA

3.1 Dados

O presente estudo refere-se à coleta dos EIAs de 25 PCHs do estado de Mato Grosso. Para cada uma das PCHs, foram obtidos os EIAs completos em formato PDF, disponíveis no *site* da SEMA-MT. Além disso, a localização espacial de cada PCH foi determinada pela localização da barragem e posteriormente organizada em forma de mapa utilizando o *software* ArcGIS 10.8 (ESRI, 2020).

3.2 Preparação dos dados

A primeira etapa do estudo envolveu a preparação dos dados textuais. Foram listados todos os arquivos de texto na pasta especificada, e seus nomes foram extraídos. Cada arquivo foi lido e transformado em um único texto completo, sendo então armazenado em um *data frame* com colunas para o ID e o texto correspondente. Essa organização inicial dos dados permitiu um manuseio eficiente dos textos e facilitou as etapas subsequentes de análise.

3.3 Pré-processamento

O pré-processamento dos textos foi uma etapa crucial para garantir a qualidade e a consistência dos dados analisados. Primeiramente, foi realizada a lematização das palavras, reduzindo-as à sua forma básica (lemmas). Esse processo foi realizado utilizando um modelo pré-treinado UDPipe para o idioma Português, por meio do pacote udpipe (Wijffels, 2023). A lematização é essencial para unificar diferentes formas de uma palavra, permitindo uma análise mais precisa dos textos.

Em seguida, os textos passaram por um rigoroso processo de limpeza e normalização. Este processo envolveu a conversão das palavras para

letras minúsculas e a remoção de elementos irrelevantes, como palavras de pouca significância (*stopwords*), pontuação, números e hífens. Para isso, foram utilizados os pacotes *stopwords* (Benoit; Muhr; Watanabe, 2021) e tm (Feinerer; Hornik; Meyer, 2008). Além disso, o texto foi processado para remover espaços em branco excessivos e garantir a uniformidade do *corpus*. Este pré-processamento é fundamental para otimizar as etapas subsequentes de análise, eliminando ruídos e facilitando a extração de informações significativas.

3.4 Exploração dos dados: nuvens de palavras, bigramas e grafos

Com o *corpus* preparado, foram geradas nuvens de palavras para visualizar as mais frequentes nos textos, facilitando a identificação rápida dos temas mais recorrentes nos EIAs das PCHs. Utilizando *software R Core Team* (2023), essas nuvens destacam as palavras mais usadas de forma visualmente intuitiva (Zizka; Darena; Svoboda, 2019).

Além disso, foram criadas nuvens de bigramas, que combinam duas palavras frequentemente associadas, como "impacto ambiental" ou "gestão hídrica". Isso permitiu uma análise mais detalhada, identificando expressões-chave que enriquecem a compreensão dos temas abordados.

A análise foi complementada por grafos, que mapeiam as relações entre palavras e bigramas, revelando como os termos se interligam. Essa abordagem destacou padrões de coocorrência e *clusters* de palavras que indicam tópicos importantes e diferentes focos temáticos nos estudos (Zizka; Darena; Svoboda, 2019).

Essas análises combinadas, nuvens de palavras, nuvens de bigramas e grafos — forneceram uma base sólida para a interpretação dos EIAs das PCHs, permitindo uma visualização clara dos principais tópicos e suas inter-relações nos documentos analisados.

4 RESULTADOS E DISCUSSÕES

4.1 Descrição dos Estudos de Impacto Ambiental

O presente estudo baseou-se na análise de 25 EIA de PCHs no estado de Mato Grosso, disponíveis no *site* da SEMA-MT. No entanto, devido a problemas como *links* de acesso indisponíveis, documentos em formato de fotocópia ilegível ou ausência de capítulos essenciais, 17 desses estudos não puderam ser utilizados. As exclusões ocorreram nas seguintes PCHs:

- PCHs Sumidouro, Perdidos, Guapira II, Iratambé I e II, Angatu I e II e Perudá: *link* de acesso indisponível.
- PCHs Bocaiuva, Foz do Cedro, Rio Claro, São Lourenço e Sauê: estudos fotocopiados, impossibilitando a leitura pelos *softwares* utilizados.
- PCHs Jesuíta, Guaporé, Saracura e Jacutinga: volume ou capítulo referente aos impactos ambientais não estavam redigidos ou não houve *upload*.

Os 15 EIAs utilizados neste trabalho referem-se a um total de 25 PCHs, organizadas conforme o Quadro 1 e com sua localização destacada na Figura 1. A figura apresenta a distribuição das PCHs no estado de Mato Grosso, cujos EIAs estão disponíveis no *site* da SEMA-MT. No entanto, dos 25 EIAs, 17 não puderam ser aproveitados, sendo representados no mapa em preto, devido a problemas como *links* inacessíveis ou documentos ilegíveis. As PCHs efetivamente analisadas estão indicadas em amarelo.

A figura também evidencia a distribuição das PCHs nas bacias hidrográficas Amazônica (verde), Tocantins-Araguaia (azul) e Paraguai (laranja), que possuem características hidrológicas distintas, influenciando os impactos ambientais dessas usinas. As PCHs representadas em preto são aquelas cujos EIAs estavam disponíveis, mas não puderam ser utilizados, enquanto as PCHs em amarelo foram analisadas.

A concentração maior de PCHs nas áreas azul e verde, que correspondem às bacias do Tocantins-Araguaia e Amazônica, respectivamente, ressalta a relevância dessas diferenças geográficas e hidrológicas para os impactos ambientais de cada usina. Dessa forma, a figura não apenas contextualiza a localização das PCHs estudadas, mas também destaca a importância das

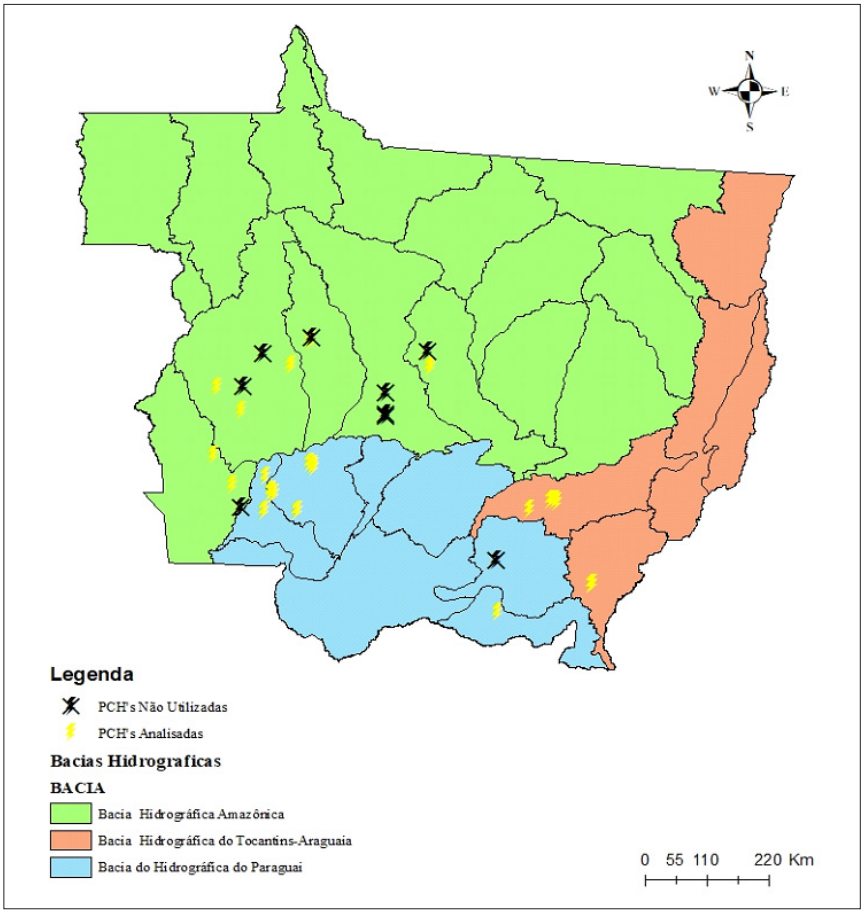
bacias hidrográficas para a análise dos impactos ambientais no estado de Mato Grosso.

Quadro 1 – Organização dos estudos de impacto ambiental por PCH

EIA	PCHs
EIA 1	Barra da Onça, Alto Garças (MLT, 2016)
EIA 2	Cabaçal I, Cabaçal IV, Cabaçal V, Cabaçal VI, Cabaçal VII, Cabaçal VIII (PROGEPLAN, 2021)
EIA 3	Cristalina (PROGEPLAN, 2021)
EIA 4	Cumbuco e Geóloga Lucimar Gomes (MRS, 2020a)
EIA 5	Entre Rios (MRS, 2020b)
EIA 6	Estivadinho (PROAMB, 2017)
EIA 7	Formoso I, Formoso II, Formoso III (SEIVA, 2011)
EIA 8	Galera (CTE, 2013)
EIA 9	Itiquira III (CTE, 2012)
EIA 10	Juína (PROAMB, 2013)
EIA 11	Mogno (Energia Consult, 2012)
EIA 12	Octacilio Lucion (MRS, 2021)
EIA 13	Rancho Grande, Progresso (BOVEN Energia, 2020)
EIA 14	Sacre 14 (PROAMB, 2016)
EIA 15	Vila União (MRS, 2020)

Fonte: EIAs disponíveis no *site* da SEMA-MT. Elaboração: autores 2024.

Figura 1 – Mapa de localização das PCHs utilizadas e não utilizadas na análise



Fonte: estudos de impactos ambientais disponíveis no *site* da SEMA-MT. Elaboração: autores 2024.

Esses 15 EIAs selecionados oferecem uma base robusta para a análise dos impactos ambientais associados às PCHs no estado de Mato Grosso. Inicialmente, foi realizada uma análise visual do conteúdo desses estudos para verificar sua qualidade e completude, o que revelou uma estruturação textual comum na maioria dos EIAs. Cada um dos estudos foi minuciosamente examinado para identificar os principais temas e preocupações ambientais, com foco nos impactos descritos.

Em sua maioria, os estudos de impacto ambiental seguem a seguinte estruturação:

- Apresentação e caracterização do empreendimento: informações técnicas detalhadas sobre as PCHs, como capacidade instalada, altura da queda d'água, tipo de barragem e equipamentos utilizados.
- Legislações pertinentes: destaque para as legislações aplicáveis desde a fase de implantação até a operação do empreendimento, assegurando o cumprimento das normas ambientais vigentes.
- Objetivos e justificativa do empreendimento: definição dos objetivos do empreendimento e justificativas para sua instalação, geralmente relacionadas ao desenvolvimento econômico e energético da região.
- Localização: informações detalhadas sobre o rio onde a PCH será instalada, o município envolvido e as vias de acesso.
- Definição das áreas de influência: descrição das áreas de influência direta, indireta e total do empreendimento, fundamentais para a avaliação dos impactos ambientais.
- Diagnósticos ambientais: diagnósticos físicos, bióticos, geológicos e socioeconômicos, fornecendo uma visão abrangente do estado ambiental da área antes da implementação do projeto.
- Impactos ambientais e programas de recuperação: detalhamento dos impactos ambientais previstos e dos programas de recuperação ambiental para mitigar esses impactos.
- Equipe técnica: identificação da equipe responsável pelo empreendimento e pelos estudos realizados durante o EIA, garantindo transparência e credibilidade.
- Considerações finais: resumo das conclusões e recomendações para a implementação do empreendimento de forma ambientalmente responsável.

A estruturação textual dos EIAs é obrigatória para todos os empreendimentos que necessitam de um estudo de impacto ambiental, conforme a Resolução CONAMA n. 01/1986 (Brasil, 1986). O artigo 5º, incisos II, III e IV, define os requisitos para a identificação, avaliação e definição das áreas de influência dos impactos, enquanto o artigo 6º especifica as atividades técnicas mínimas que devem ser incluídas no EIA. Os EIAs analisados seguem essas

exigências, cobrindo os aspectos técnicos e ambientais necessários para decisões informadas sobre a viabilidade e mitigação dos impactos. Embora a ordem dos estudos possa variar conforme a empresa responsável, todos os EIAs atendem às normas legais, garantindo informações consistentes e comparáveis para a análise ambiental das PCHs.

5 MINERAÇÃO DE TEXTO

A Tabela 1 apresenta o número de palavras antes e após o pré-processamento dos textos dos EIAs, enquanto a Tabela 2 exibe os valores dos índices de diversidade léxica, como Simpson e Berger-Parker, tanto antes quanto depois do pré-processamento.

Tabela 1 – Número de palavras antes e após o pré-processamento

EIA	Antes	Após	Percentual de redução
EIA 1	116682	23489	79,87%
EIA 2	419405	51877	87,63%
EIA 3	324585	40441	87,54%
EIA 4	182224	9824	94,61%
EIA 5	184779	9881	94,65%
EIA 6	21596	4316	80,01%
EIA 7	120151	14267	88,13%
EIA 8	61347	9394	84,69%
EIA 9	77469	8886	88,53%
EIA 10	49208	14665	70,20%
EIA 11	15072	2323	84,59%
EIA 12	62398	9315	85,07%
EIA 13	156314	19492	87,53%
EIA 14	11869	2119	82,15%
EIA 15	193450	11070	94,28%

Fonte: estudos de impactos ambientais disponíveis no *site* da SEMA-MT. Elaboração: autores 2024.

O pré-processamento dos textos dos EIAs foi essencial para preparar os dados para análises posteriores, com uma redução significativa do número

de palavras entre 70,20% e 94,65%. Esse processo envolveu a remoção de palavras irrelevantes (como *stopwords*), pontuação, números e a normalização do texto (conversão para minúsculas), garantindo uma análise mais limpa e focada.

Apesar da redução substancial de palavras, a diversidade léxica dos textos, medida pelos índices de Simpson e Berger-Parker, mostrou que essa simplificação não afetou drasticamente a riqueza dos textos. O índice de Simpson aumentou em todos os EIAs, sugerindo a simplificação do texto. Após o pré-processamento, os valores do índice de Berger-Parker diminuíram, indicando que o pré-processamento reduziu a dominância de termos comuns e repetitivos, promovendo uma distribuição mais equilibrada das palavras restantes.

Tabela 2 – Índices de Diversidade Lexical: Simpson e Berger-Parker Antes e Após a Intervenção

Simpson		Berger Parker	
Antes	Apos	Antes	Apos
0,9904	0,9968	0,0689	0,0303
0,9893	0,9967	0,0724	0,0352
0,9892	0,9944	0,0728	0,0418
0,9890	0,9958	0,0732	0,0272
0,9874	0,9964	0,0859	0,0281
0,9856	0,9950	0,0945	0,0269
0,9902	0,9963	0,0664	0,0240
0,9902	0,9969	0,0654	0,0247
0,9890	0,9969	0,0700	0,0251
0,9896	0,9963	0,0710	0,0236
0,9895	0,9964	0,0723	0,0237
0,9890	0,9960	0,0773	0,0259
0,9904	0,9981	0,0700	0,0198
0,9890	0,9966	0,0754	0,0264
0,9899	0,9971	0,0681	0,0215

Fonte: estudos de impactos ambientais disponíveis no *site* da SEMA-MT. Elaboração: autores 2024.

Estudos anteriores, como os de Suhaidi, Kadir e Tiun (2023) e Uysal e Gunal (2014), também confirmam que o pré-processamento é fundamental para melhorar a qualidade e a precisão das análises textuais. Ao reduzir o volume de dados e destacar termos mais relevantes, essa etapa foi crucial para a aplicação de técnicas de mineração de texto, como a análise de grafos e outros métodos de representação textual.

Além disso, o pré-processamento reduziu o espaço vetorial multidimensional dos textos, fundamental para melhorar a eficiência das análises subsequentes, conforme discutido por Uysal e Gunal (2014). Com o tratamento adequado, as informações essenciais foram preservadas, proporcionando uma base sólida para a aplicação de métodos mais avançados.

O pré-processamento não só simplificou os dados como garantiu que as análises fossem realizadas com foco nas informações mais relevantes, permitindo uma compreensão mais profunda dos padrões e conexões semânticas presentes nos EIAs. Esses resultados estão em linha com as práticas recomendadas na literatura de mineração de texto, reforçando a importância de um pré-processamento bem executado para a análise detalhada de grandes volumes de texto.

5.1 Frequência de palavras

A análise de frequência de palavras (Figura 2) destaca as repetições de termos dentro do texto, com as palavras mais frequentes ganhando maior destaque no centro da nuvem. Esse tipo de visualização permite identificar rapidamente os principais temas e conceitos discutidos, sendo uma ferramenta eficaz na análise de conteúdo e mineração de texto (Zizka; Darena; Svoboda, 2019).

Figura 2 – Nuvem de palavras de 15 estudos de impactos ambientais utilizados, referentes a impactos ambientais de um total de 25 PCHs

Fonte: estudos de impactos ambientais disponíveis no *site* da SEMA-MT. Elaboração: autores, 2024.

Os principais temas abordados na nuvem estão relacionados aos IMPACTOS AMBIENTAIS de EMPREENDIMENTOS, com palavras como IMPACTO, EMPREENDIMENTO, PROGRAMA, ÁREA e AMBIENTAL em destaque. Esses termos indicam um foco em estudos de Impacto Ambiental (EIA/RIMA), essenciais para avaliar os efeitos de projetos sobre o meio ambiente. Palavras como ALTERAÇÃO, FAUNA, VEGETAÇÃO e RESERVATÓRIO reforçam a preocupação com mudanças nos ecossistemas e a necessidade de medidas mitigadoras (Jesus *et al.*, 2018).

A presença de termos como MONITORAMENTO, MEDIDA e PROBABILIDADE destaca a importância das ações de controle de impactos. Essas medidas, discutidas amplamente na literatura de gestão ambiental, são fundamentais para garantir que os danos ambientais sejam mitigados. Assim, a nuvem de palavras identifica os pontos críticos dos textos, enfatizando a relevância da gestão ambiental nos empreendimentos analisados (Sánchez, 2013).

5.2 Frequência de Bigramas

Na Figura 3, é apresentada a nuvem de palavras dos bigramas, uma ferramenta importante para identificar e analisar as relações entre palavras que ocorrem em sequência. A análise por bigramas permite não apenas detectar os termos mais frequentes, mas também compreender melhor os contextos em que eles aparecem juntos, revelando padrões relevantes em textos sobre IMPACTOS AMBIENTAIS. Neste caso, o bigrama IMPACTO AMBIENTAL é o mais destacado, refletindo o foco dos estudos de Impacto Ambiental (EIA). Esse termo, que é central nas discussões sobre os efeitos dos EMPREENDIMENTOS no meio ambiente, aparece repetidamente porque resume as preocupações principais desses estudos, que visam avaliar as consequências ambientais de grandes projetos de infraestrutura (Sánchez, 2013).

Figura 3 – Nuvem de palavras dos bigramas de 15 estudos de impactos ambientais utilizados, referentes a impactos ambientais de um total de 25 PCHs

Fonte: estudos de impactos ambientais disponíveis no *site* da SEMA-MT. Elaboração: autores, 2024.

Além disso, bigramas como QUALIDADE DA ÁGUA evidenciam tópicos recorrentes nos EIAs, mostrando que a preservação e o monitoramento dos RECURSOS HÍDRICOS são frequentemente tratados como questões críticas nos relatórios ambientais. A ênfase nesses termos reflete a importância da água como um recurso vital, cujo uso e impacto são discutidos exaustivamente nos documentos de licenciamento ambiental. Palavras como IMPLANTAÇÃO DO EMPREENDIMENTO e LICENCIAMENTO AMBIENTAL destacam o foco nos aspectos legais e operacionais dos projetos, desde a

fase de planejamento até a execução, abordando as exigências legais e as precauções necessárias para mitigar os impactos negativos (Jesus *et al*., 2018). Esses bigramas mostram que os textos analisados estão concentrados não apenas nos efeitos ambientais, mas também nas responsabilidades de implementação sustentável.

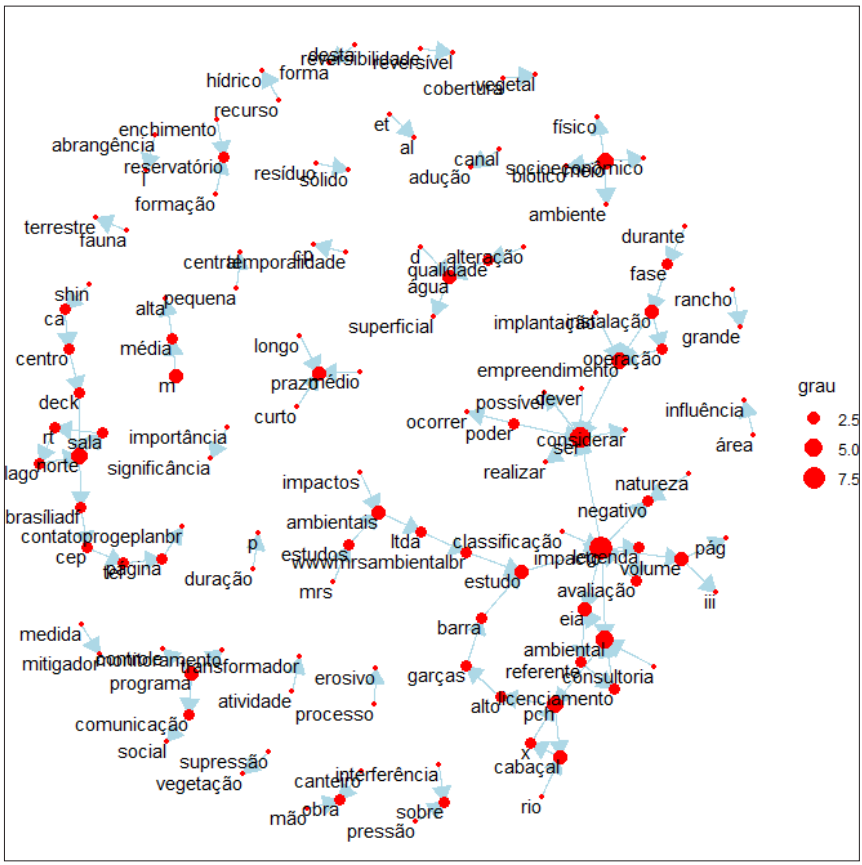
Por outro lado, a análise de bigramas também revela algumas limitações técnicas, como a presença de ENDEREÇOS ELETRÔNICOS E FÍSICOS. Estes são elementos que não estão diretamente relacionados ao conteúdo principal dos estudos de impacto, mas acabam aparecendo devido à análise de rodapés, cabeçalhos ou seções de contato presentes nos documentos. A inclusão desses termos menos relevantes ressalta a necessidade de um pré-processamento mais cuidadoso dos textos, eliminando informações supérfluas que possam "poluir" os resultados da análise. Esse pré-processamento, ao remover detalhes técnicos ou estruturais, pode garantir que a mineração de texto foque exclusivamente no conteúdo semântico significativo, permitindo uma análise mais precisa e focada nos temas centrais dos estudos (Ur-Rahman, 2017).

5.3 Grafos de palavras

A análise de grafos permite entender as conexões entre palavras em um texto por meio da coocorrência. As palavras mais frequentes tornam-se os vértices do grafo, enquanto as arestas representam as ligações entre elas, demonstrando como os termos se conectam em um contexto específico. Na Figura 4, visualizamos essas relações, em que as palavras mais centrais e com maior ocorrência se destacam.

Conforme explicado por Dutt-Ross e Cruz (2021), o algoritmo utilizado na análise aplica uma função decrescente à distância entre pares de palavras, atribuindo maior pontuação a termos que aparecem frequentemente juntos. Isso faz com que palavras que coocorrem regularmente tenham uma forte influência na estrutura do grafo, enquanto palavras que aparecem mais esporadicamente também contribuem, mas de forma mais sutil.

Figura 4 – Grafo de palavras considerando a ocorrência de pelo menos 100 vezes em 15 estudos de impactos ambientais utilizados, referindo-se a impactos ambientais de um total de 25 PCHs



Fonte: estudos de impactos ambientais disponíveis no *site* da SEMA-MT. Elaboração: autores 2024.

No grafo apresentado, é possível perceber como essas associações funcionam. A palavra MEIO está altamente conectada a termos como FÍSICO, BIÓTICO, SOCIOECONÔMICO e AMBIENTE, sugerindo que esses conceitos são frequentemente discutidos em conjunto, provavelmente no contexto das descrições ambientais dentro dos EIAs. Além disso, palavras como IMPACTO e SER apresentam alta coocorrência, o que reforça a centralidade

do tema “impacto ambiental” e a presença frequente do verbo “ser” nas construções textuais.

Os vértices do grafo variam em tamanho conforme o grau de coocorrência: vértices maiores indicam palavras com maior grau de conexão, ou seja, termos que aparecem frequentemente associados a outros. A legenda mostra diferentes graus de conexão, e os vértices maiores indicam maior importância nas interações de coocorrência.

Algumas palavras-chave se destacam no grafo pela centralidade e número de conexões. Termos como IMPACTOS, AMBIENTAIS, IMPLANTAÇÃO, EMPREENDIMENTO, e QUALIDADE DA ÁGUA são fortemente conectados, sugerindo que esses conceitos são centrais e recorrentes nos textos analisados. Além disso, agrupamentos de palavras como MEIO, FÍSICO, BIÓTICO e SOCIOECONÔMICO sugerem uma inter-relação entre as diferentes dimensões do meio ambiente discutidas nos EIAs. Termos como DURAÇÃO, PRAZO, MITIGAÇÃO e MONITORAMENTO também aparecem com uma quantidade significativa de conexões, indicando sua relevância no contexto da avaliação de impactos ambientais e no planejamento de medidas mitigadoras.

Assim, o grafo facilita a identificação das relações semânticas mais fortes nos textos, proporcionando uma visão clara das interconexões entre os principais conceitos discutidos, além de destacar as palavras mais importantes e como elas interagem dentro do contexto dos EIAs.

6 CONSIDERAÇÕES FINAIS

A aplicação de técnicas de mineração de texto na análise dos EIA de PCHs em Mato Grosso demonstrou ser uma abordagem eficaz para extrair informações valiosas de grandes volumes de dados textuais. Ferramentas como nuvens de palavras, bigramas e grafos permitiram identificar temas recorrentes e inter-relações importantes nos documentos, destacando conceitos centrais como impacto ambiental, qualidade da água e medidas de mitigação. Essa análise não apenas aprofundou a compreensão dos impactos ambientais associados aos empreendimentos, mas também forneceu subsídios para o desenvolvimento de políticas públicas e estratégias de gestão ambiental mais eficazes.

Durante a análise dos EIAs de 25 PCHs disponíveis no *site* da SEMA-MT, vários desafios foram identificados. Dos 25 estudos, 17 não puderam ser utilizados devido a problemas de acesso, baixa qualidade dos documentos ou ausência de informações cruciais. Esse cenário ressalta a necessidade urgente de uma melhor organização e acessibilidade dos documentos disponibilizados. Dos 15 estudos analisados, a aplicação de técnicas de mineração de texto revelou uma redução significativa no número de palavras após o pré-processamento, que variou de 70,20% a 94,65%, facilitando a identificação de padrões semânticos e focos temáticos como SER, IMPACTO, EMPREENDIMENTO, AMBIENTAL e PROGRAMA, os quais predominam nos relatórios.

Além das questões técnicas, o estudo indicou a necessidade de uma revisão das legislações ambientais que regulamentam os EIAs, como a Resolução CONAMA n. 01. de 1986, que, diante dos avanços tecnológicos e das novas demandas, encontra-se ultrapassada. Também é essencial melhorar a organização dos documentos no *site* da SEMA-MT, com a disponibilização de EIAs completos e bem categorizados, facilitando o acesso e a consulta pública. A falta de padronização entre os estudos contratados por diferentes empresas também evidenciou a necessidade de unificar a apresentação e estrutura dos EIAs, promovendo uma maior transparência e eficiência na análise dos impactos ambientais.

REFERÊNCIAS

AYKURT, A. Y.; SESEN, E. Social media in social organization. *European Scientific Journal*, [*S. l.*], v. 13, n. 20, p. 1, 2017. DOI: 10.19044/esj.2017.v13n20p1.

BENOIT, K.; MUHR, D.; WATANABE, K. Multilingual stopword lists. *CRAN*, [*S. l.*], 2021.

BOVEN ENERGIA. *EIA do Complexo Hidrelétrico Rancho Grande Volume III*. São Paulo: BOVEN Energia, 2020.

BRASIL. Ministério do Meio Ambiente [MMA]. Conselho Nacional do Meio Ambiente [CONAMA]. *Resolução CONAMA n. 237*, de 19 de dezembro de 1997 Publicada no DOU no 247, de 22 de dezembro de 1997, Seção 1, páginas 30841-30843. Brasília, DF: MMA; CONAMA, 1997.

BRASIL. *Decreto no 99.274*, de 6 de junho de 1990. Regulamenta a Lei nº 6.902, de 27 de abril de 1981, e a Lei nº 6.938, de 31 de agosto de 1981, que dispõem, respectivamente sobre a criação de Estações Ecológicas e Áreas de Proteção Ambiental e sobre a Política Nacional do Meio Ambiente, e dá outras providências. Brasília, DF: Presidência da República; Casa Civil; Subchefia para Assuntos Jurídicos, 1990.

BRASIL. Ministério do Meio Ambiente [MMA]. Conselho Nacional do Meio Ambiente [CONAMA]. *Resolução CONAMA n. 001*, de 23 de janeiro de 1986. Dispõe sobre critérios básicos e diretrizes gerais para a avaliação de impacto ambiental. Brasília, DF: MMA; CONAMA, 1986.

BULEGON, H.; MORO, C. M. C. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. *Journal of Health Informatics*, [*S. l.*], v. 2, n. 2, 2010.

CHOWDHURY, G. G. Natural language processing. *Annual Review of Information Science and Technology*, [*S. l.*], v. 37, n. 1, p. 51-89, 31 jan. 2005.

CTE. *Estudo de Impacto Ambiental da PCH Galera Volume III.* Goiânia: Centro Tecnológico De Engenharia ltda, 2013.

CTE. *Estudo de Impacto Ambiental da PCH Itiquira III Volume III*. Goiânia: Centro Tecnológico De Engenharia *ltda*, 2012.

DUTT-ROSS, S.; CRUZ, B. P. A. Análise Quantitativa de Textos: apresentação e operacionalização da técnica via Twitter. *Administração: Ensino e Pesquisa*, Rio de Janeiro, v. 22, n. 1, jan./abr. 2021. DOI: https://doi.org/10.13058/raep.2021.v22n1.1859.

ENERGIA CONSULT. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Mogno.* São Paulo: Energia Consult Engenharia, Consultoria e Gerenciamento de Projetos Ltda, 2012.

ESRI. *ArcGIS 10.8*: geographic information system software. Redlands: Environmental Systems Research Institute, 2020.

FEINERER I.; HORNIK K.; MEYER, D. Text Mining Infrastructure in R. *Journal of Statistical Software*, [*S. l.*], v. 25, n. 5, mar. 2008.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook*: advanced approaches in analyzing unstructured data. Cambridge: University Press, 2007.

JESUS, E. N.; FEITOSA, F. R. S.; SOBRAL, I. S.; SILVA, H. P. Educação ambiental e o licenciamento: um olhar sobre os relatórios de impactos ambientais do estado de Sergipe. *Revista de Ciências Ambientais*, Canoas, v. 12, n. 1, 2018. DOI: https://doi.org/10.18316/rca.v12i1.3449

MLT. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Barra da Onça e Alto Garças*. Belo Horizonte: MLT Engenharia de Projetos Ambientais, 2016.

MOHAMMAD, S. M.; TURNEY, P. D. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, [S. l.], v. 29, n. 3, p. 436-65, 2013.

MRS. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Otacílio Lucion*. Brasília: MRS Ambiental, 2021.

MRS. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH's Cumbuco e Geóloga Lucimar Gomes*. Brasília: MRS Ambiental, 2020a.

MRS. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Entre Rios*. Brasília: MRS Ambiental, 2020b.

MRS. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Vila União*. Brasília: MRS Ambiental, 2020c.

NAVEGA, S. Princípios essenciais do data mining. *Intelliwise Research and Training*, São Paulo, ago. 2002.

PROAMB*. Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Estivadinho III*. Bento Gonçalves: PROAMB, 2017

PROAMB. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Sacre 14. Bento Gonçalves: PROAMB Projetos e Consultoria Ambiental ltda, 2016. 385 p.* Disponível em: http://sema.mt.gov.br/eia-rima. Acesso em: 12 jul.

PROAMB. *Estudo de Impacto Ambiental de Pequenas Centrais Hidrelétricas PCH Juina 117.* Bento Gonçalves: PROAMB, 2013.

PROGEPLAN. *Estudo de Impacto Ambiental (EIA) Volume III – Análise e Avaliação dos Impactos Ambientais*; Medidas Mitigadoras Preventivas e Compensatórias; Programas Ambientais; Alternativas Locacionais; Prognóstico Ambiental e Análise de Risco. Brasília: PROGEPLAN, 2021.

R CORE TEAM. R: a language and environment for statistical computing. *R*

Foundation for Statistical Computing, Vienna, 2023.

SÁNCHEZ, L. E. *Avaliação de impacto ambiental*: conceitos e métodos. 2. ed. São Paulo: Oficina de Textos, 2013. 87p.

SEIVA. *Estudo de Impacto Ambiental das pequenas centrais hidrelétricas formoso I, II e III, Volume VI, Capitulo VII, Cuiabá-MT*: Seiva, 2011. 90 p.

SRIVASTAVA, A. N.; SAHAMI, M. (Ed.). *Text Mining*: classification, clustering, and applications. Boca Raton: Chapman & Hall/CRC, 2009.

SUHAIDI, M.; KADIR, R. A.; TIUN, S. The impact of preprocessing techniques towards word embedding. *In*: BADIOZE ZAMAN, H. *et al*. (Ed.). *Advances in Visual Informatics*: IVIC 2023. Singapore: Springer, 2023. DOI: https://doi.org/10.1007/978-981-99-7339-2_35

TURBAN, E.; WETHERBE, J. C.; MCLEAN, E.; LEIDNER, D. E. *Tecnologia da informação para gestão*: transformando os negócios na economia digital. 6. ed. Porto Alegre: Bookman, 2010. 680 p.

UR-RAHMAN, N. Textual data mining for knowledge discovery and data classification: a comparative study. *European Scientific Journal*, [*s. l.*], v. 13, n. 21, p. 429-53, 2017. DOI: http://dx.doi.org/10.19044/esj.2017.v13n21p429.

UYSAL, A. K.; GUNAL, S. The impact of preprocessing on text classification. *Information Processing & Management*, [*S. l.*], v. 50, n. 1, p. 104-112, 2014. ISSN 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2013.08.006.

WIJFFELS, J. UDPipe Natural Language Processing – Basic Analytical Use Cases. *CRAN*, [*S. l.*], 2023.

ŽIŽKA, J.; DAŘENA, F.; SVOBODA, A. *Text mining with machine learning*: principles and techniques. Boca Raton: CRC Press, 2019.