A mineração de textos em um mundo de dados abundantes

Caroline Mengue Bendl[1]
Luan Tavares dos Santos[2]
Luiz Gustavo da Silva Moraes[3]
Mikael Medeiros Ramos [4]
Clóvis Silveira[5]

Resumo: O volume, acessibilidade e importância dos dados não estruturados na forma de textual e digital estão aumentando rapidamente para pesquisas sobre inovação, enquanto abordagens tradicionais para análise de texto são limitadas para processar uma grande quantidade de dados. A mineração de texto é uma extensão da área de mineração de dados e pode ser definida como um processo de extração de informações desconhecidas e úteis de arquivos textuais, escritos em linguagem natural. Para que isso se torne possível, são necessários vários passos como extração dos dados, processamento e análise. Este trabalho tem por objetivo geral aprofundar a compreensão da mineração de textos no contexto da ciência de dados, explorando suas definições, metodologias e aplicações. Além disso, apresenta um estudo qualitativo na área, onde é feita a análise dos resultados da pesquisa feita com base em dez artigos acadêmicos a fim de mineração. Resultados apontam que foi possível realizar a mineração de textos em um corpus de 10 artigos e descobrir palavras mais utilizadas para descoberta de conhecimentos.

Palavras-chave: Mineração de Textos, Ciência de Dados, Mineração de Dados.

1 INTRODUÇÃO

Na era digital em que vivemos, a criação de informações textuais é exponencial. Encontramo-nos em uma vastidão de dados não estruturados, como documentos, arquivos PDF, páginas Web, posts. No entanto, tais dados representam um grande desafio para a extração de informações valiosas e conhecimento significativo.

A maior parte das informações disponíveis não está armazenada de forma

[1] Discente do Curso de Graduação em Ciência da Computação do Centro Universitário Cesuca. E-mail: carolinebendl.cb@gmail.com
[2] Discente do Curso de Graduação em Ciência da Computação do Centro Universitário Cesuca. E-mail: Luantasan@gmail.com
[3] Discente do Curso de Graduação em Ciência da Computação do Centro Universitário Cesuca. E-mail: Luizgustavo.041200@gmail.com
[4] Discente do Curso de Graduação em Ciência da Computação do Centro Universitário Cesuca. E-mail: mikael10_mr@hotmail.com
[5] Docente do curso de Ciência da Computação e Análise e Desenvolvimento de Sistemas no Centro Universitário Cesuca. Doutor em Informática na Educação. E-mail: clovis.silveira@cesuca.edu.br

estruturada, e sim o oposto, os dados se encontram digitalmente, de forma textual: livros, jornais, e outras infinitas fontes de dados não estruturados. Essa situação incentivou a criação de uma subárea na mineração de dados e ficou conhecida como mineração de texto, que tem como objetivo encontrar termos significativos em documentos detexto contendo muitas informações e criar padrões e conexões entre eles com base na frequência e na natureza temática dos termos encontrados (Serapião, 2010).

A mineração de textos é capaz de compreender a linguagem natural dos arquivos textuais e consegue lidar com sua imprecisão. Essa área engloba certas áreas da informática, como *machine learning,* recuperação de informação, estatística e linguagem computacional, necessárias para atransformação dos textos em algo computacionalmente compreensível (Machado *et al*., 2010).

Este trabalho tem como objetivo geral proporcionar uma ampla visão no mundo da mineração de textos. Para tanto, será apresentada a conexão dessa área com a ciência de dados, suas definições, técnicas utilizadas e suas aplicações. Este conta, também, com um breve estudo realizado na área, onde dez artigos coletados no Google Acadêmico compuseram o Corpus da pesquisa a fim de concluir as etapas da mineração textual.

O presente trabalho está dividido em 7 seções ao total, contendo, em primeiro lugar, a definição da ciência de dados e sua conexão com a mineração de textos. A seguir, é dissertado sobre a mineração de dados textuais, trazendo a definição da Descoberta de Conhecimento a partir de dadosnão estruturados e o Processamento de Linguagem Natural. Esse item é seguido das etapas da mineração, onde se fala sobre os passos necessários para uma extração eficiente. Em seguida, é apresentado as metodologias usadas no estudo e o próximo item mostra os resultados obtidos.

2 CIÊNCIA DE DADOS

O termo Data Science surgiu em meados dos anos 60, com o intuito de referir-se à ciência de dados como um todo, e por ser considerada uma ciência nova, às vezes pode ser mal interpretada. Como citam Cielen, Meysman e Ali (2016) a ciência de dados é uma extensão crescente da estatística que lida com uma enorme quantidade de dados. Os autores também fazem menção às descrições de vagas para cientistas de dados e um dos principais requisitos é a habilidade de trabalhar

com Big Data, além de possuir experiência com o desenvolvimento de algoritmos e machine learning.

A ciência de dados atua em um amplo escopo, sendo muito aplicada em anúncios erecomendações para um determinado mercado-alvo, que se baseia no comportamento do cliente. Diversas empresas utilizam da ciência de dados para se destacar no mercado, tais como Amazon, Netflix, entre muitas outras, e devido ao uso dessas estratégias, muitas empresas evoluem para empresas de mineração de dados (Provost; Fawcett, 2013).

Ao passar dos anos, é notável o crescimento na quantidade de material textual que se dispõenas redes, que leva a novas descobertas e cria novos caminhos para a pesquisa. Nesse contexto, para Hassani *et al*. (2020), a mineração de texto obteve interesse nesse campo de rápido desenvolvimento de abordagens analíticas de Big Data, com o uso de processamento paralelo, deep learning e reconhecimento de padrões com dados textuais dado a crescente importância da inteligência artificiale sua implementação em plataformas digitais. O texto está se tornando cada vez mais necessário para todos os modelos de negócio, pesquisas de mercado, estratégias de marketing e até em tomada de decisões.

3 MINERAÇÃO DE TEXTOS

Ao longo de muitos anos e em vários campos científicos, a mineração de texto tornou- se um corpo de literatura no ramo da análise de dados textual. Isso inclui campos como estatísticas, ciênciasda computação, linguística computacional e biblioteconomia. Com isso, terminologias como mineração de texto, análise computacional de conteúdo e processamento de linguagem natural evoluíram (Antons, 2020).

No âmbito de mineração de textos, é necessário que, anteriormente, se aborde o assunto de Recuperação de Informações, onde é selecionado, a partir do acervo, documentos de tópicos específicos à escolha do usuário. Esse processo identifica quais informações são relevantes à necessidade do usuário, a partir de uma coleção de documentos, determinada Corpus.

A necessidade da extração de informações em bases de dados não estruturadas fez com que os Sistemas de Recuperação de Informações (SRI 's) fossem criados, estes se baseiam na busca por similaridade e palavra-chave.

De acordo com Morais e Ambrósio (2007), a mineração de textos é uma

evolução da área de Recuperação de Informações, que analisa e extrai dados a partir de textos, também conhecida como KDT (*Knowledge Discovery from Text*), ou Descoberta de Conhecimento a partir de dados não estruturados, diferentemente de KDD (*Knowledge Discovery in Databases*) que extrai informações debases de dados estruturadas.

A mineração de textos pode ser usada nas mais diversas áreas, como por exemplo na área damedicina, com um volume enorme de dados como prontuários, registros, etc.Também pode ser utilizada para analisar sentimentos em pesquisas ou questionários, onde é possível que o público responda de forma elaborada (Pezzini, 2016).

Sendo assim, em consequência do armazenamento de informações ser mais frequente na forma textual, o KDT teria um maior potencial do que o KDD, devido a cerca de 80% dos dados em organizações encontrarem-se em forma de texto. Mineração de textos, recuperação de informação e KDT dependem quase diretamente de Processamento de Linguagem Natural, especialmente usando processos de linguística computacional (Morais; Ambrósio, 2007).

3.1 DADOS TEXTUAIS NÃO ESTRUTURADOS

Dados não estruturados são metadados organizados de forma complexa, diferentemente do queocorre com os dados estruturados. Por exemplo: textos produzidos em Word, imagens, comentários no Facebook, entre outros dados que não possuem uma estrutura rígida ou fixa. Eles podem ser gerados a partir de várias fontes, incluindo arquivos de áudio, vídeos, imagens, postagens em redes sociais e arquivos de texto. É importante saber como gerenciar esses dados para extrair percepções valiosas deles.

Existem alguns métodos possíveis para armazenar dados não estruturados. Primeiro, eles devemser convertidos em um formato mais facilmente gerenciável, como o eXtensible Markup Language (XML), linguagem de marcação flexível e versátil utilizada para armazenar e transportar dados de forma legível, tanto por máquinas quanto por humanos. Um Content Addressable Storage System (CAS) é usado para armazenar dados não estruturados. Esse sistema armazena dados acessando seus metadados e atribuindo um nome exclusivoa cada item ou objeto armazenado nos dados. O objeto pode ser recuperado com base em seu conteúdo,não em sua localização. É importante ter uma estratégia robusta para

gerenciar e analisar esses dados, pois eles podem fornecer informações valiosas para a tomada de decisões.

3.2 PROCESSAMENTO DE LINGUAGEM NATURAL

O processamento de Linguagem Natural é uma área da computação que abrange aspectos da comunicação humana, estudando o desenvolvimento de algoritmos que analisam textos em linguagem natural, de uma forma simples, o PLN visa se comunicar em linguagem humana (Gonzalez; Lima, 2003).

Conforme os autores,

> A forma lógica codifica os possíveis sentidos de cada palavra e identifica os relacionamentos semânticos entre palavras e frases. Uma vez que os relacionamentos semânticos são determinados, alguns sentidos para as palavras tornam-se inviáveis e, assim, podem ser desconsiderados.

Sob visão linguística, as camadas do PLN estão divididas em cinco: (a) fonético ou fonológico,

(b) morfológico, (c) sintático, (d) semântico ou (e) pragmático. Cada um dos níveis possuem suas próprias características e objeções, mas cada aplicação do Processamento de Linguagem Natural podese voltar a um subconjunto dos níveis citados.

Fonético ou fonológico: a relação das palavras com os sons por elas produzidos;

Morfológico: construção de palavras a partir de unidades de significado primitivos e sua classificação em classes morfológicas;

Sintático: da relação das palavras na constituição de sentenças;

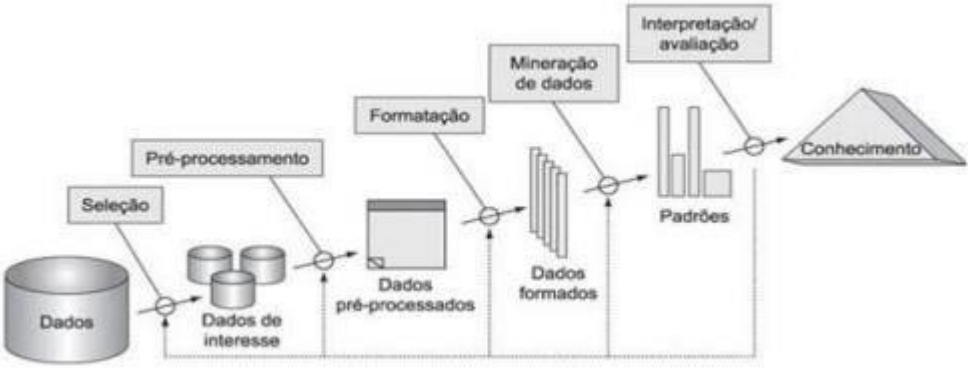
Semântico: a forma como as palavras têm suas próprias definições e se combinam para formaro significado das frases;

Pragmático: uso de frases em diferentes circunstâncias que afetam o significado.

4 ETAPAS DA MINERAÇÃO DE TEXTO

Para que se possa efetuar um processo de mineração de texto são necessárias algumas etapas,tais quais a seleção dos documentos a serem utilizados, definição do tipo de abordagem e preparação dos dados, após isso, realiza-se a análise dos resultados (Morais; Ambrósio, 2007).

Figura 1 – Uma visão geral das etapas que compõem o processo de descoberta de conhecimento em banco dedados



Fonte: Fayyad *et al*., 1996

Coleta: nessa etapa, é formada a base de documentos ou Corpus, onde os documentos contendo a base de dados não estruturada são selecionados.

Pré-processamento: consiste na preparação dos dados, visto que estes normalmente não estão em formato conveniente para a extração do conhecimento. Os dados são extraídos de bases diversas e passam por um processo de unificação, assim formando uma fonte única.

Indexação: facilita a identificação de similaridade dos significados das palavras, formando umíndice. Esse é o processo de Recuperação de Informações.

Mineração: nessa etapa, ocorrem os cálculos, inferências e a extração do conhecimento.

Análise: nessa fase, aplica-se técnicas de análise dos resultados do sistema de recuperação deinformações e gera-se os resultados do processo de mineração de textos.

5 METODOLOGIA

O estudo consistiu na pesquisa de dez artigos acadêmicos para a mineração, para que dessa forma fosse possível analisar os dados resultantes. Todos artigos foram pesquisados no Google Acadêmico, por palavras chave, sendo estas “Internet”, “Marketing” e “Strategy”.

Além disso, um filtro de data foi determinado, apenas documentos de 2012 a 2023 foram selecionados. Por último, para a composição do Corpus, todos os arquivos foram selecionados na língua inglesa, com o intuito de ampliar o escopo da busca.

No quesito de pré-processamento, todos os artigos passaram juntos por uma

ferramenta de mineração de texto, conhecida como TagCrowd. A ferramenta permite que vários arquivos sejam carregados de uma só vez, sem que seja preciso comprimir os textos em um só arquivo. Na indexação,os termos foram identificados e palavras irrelevantes foram removidas.

6 RESULTADOS DO ESTUDO

Para que o estudo pudesse ser realizado, os seguintes artigos foram selecionados para aextração do conhecimento, conforme mostra a tabela:

Tabela 1 -Corpus do Texto

Nº	Título do Artigo	Ano	Local de publicação	Nr. De palavras
1	Study on Marketing Strategy System of SMEs under Internet Background	2018	School of Business Administration, SouthChina University of Technology, Guangzhou, China	2,423
2	Evolution of Marketing Strategies: From Internet Marketing to M-Marketing	2012	Institute of Research on Population and Social Policies (IRPPS-CNR) 00198, Rome,Italy	3,962
3	Internet Marketing as a Business Necessity	2018	University of Novi Sad – Faculty of Economics in Subotica Subotica, Republicof Serbia	3,780
4	Advertising Strategy Management in Internet Marketing	2021	Dpt of Regulatory Policy Problems and Entrepreneurship Development, Instituteof Industrial Economics of the National	3,017
			Academy of Sciences in Ukraine, Kiev, Ukraine	
5	Influence of Internet Marketing Strategies on the Market Share of Online Shops in Nairobi County in Kenya	2019	University of Agriculture and Technology, P. O. Box 62000, 00200 Nairobi, Kenya	7,233
6	A Study on Internet Marketing Strategy	2014	Globus An International Journal of Management & IT A Refereed ResearchJournal	1,467

Centro Universitário Cesuca
Rua Silvério Manoel da Silva, 160, Bairro Colinas, Cachoeirinha/RS, Cep.: 94940-243 e-mail: mostracientifica@cesuca.edu.br

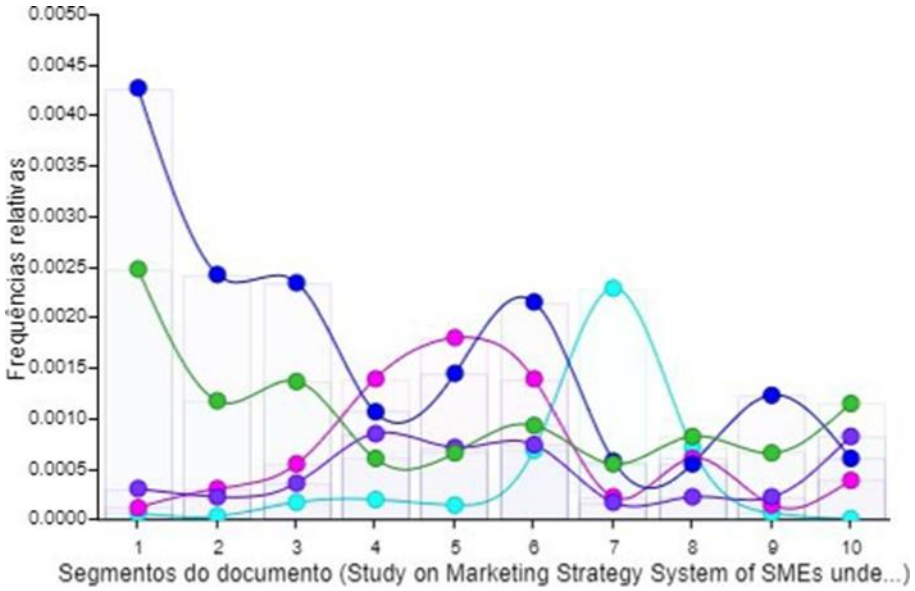
7	The Performance Audit of a Corporate Website as a Tool For ItsInternet marketing Strategy	2017	Dpt of Audit, Revision and Analysis Ternopil National Economic University 3 Peremohy sq., Ternopil, Ukraine	3,630
8	Web Site – Basic Internet MarketingStrategy Tool of Digital Companies	2019	University of Târgu Jiu, Economy Series,Issue	4,179
9	Using CRM Systems for Development and Implementation of Communication Strategies for Digital Brand Management and Internet Marketing: EU Experience	2023	JPB review	5,274
10	Modern Pricing Strategies in the Internet Market	2017	International Days of Statistics and Economics, Prague	1,666

Logo depois, após passá-los pela ferramenta TagCrowd, foi possível analisar os termos mais relevantes por artigo, e, levando esses dados em consideração, as três palavras mais frequentes no Corpus foram marketing (aparecendo 611 vezes no total), internet (em evidência 380 vezes), e online (252 vezes), como aparecem com maior destaque na figura 2:

Figura 2 – Cirrus

Assim, tendo em conta os resultados descritos na figura 2, foi possível descobrir a tendência na qual as palavras apareciam no Corpus, assim como mostra a figura 3:

Figura 3 - Tendência



7 CONSIDERAÇÕES FINAIS

Grandes quantidades de dados textuais se tornaram mais fáceis de acessar devido às novas tecnologias, que, por sua vez, acabam capturando uma porção cada vez maior de cultura, engajamento e comunicação. A aplicação de tecnologias de mineração de texto cresceu gradativamente e tornou-se muito ampla, oferecendo uma estrutura para maximizar o valor da informação contida em enormes quantidades de texto.

Como inicialmente descrito, o presente trabalho teve como objetivo trazer uma visão geral sobrea mineração de textos e fazer um breve estudo na literatura. Para alcançar tal objetivo, artigos de diferentesautores foram estudados e dissertados, assim como o uso da ferramenta TagCrowd, como previamente citada.

No segundo item, disserta-se brevemente sobre a área da ciência de dados, trazendo uma visão geral de como esse campo se relaciona com a mineração de textos.

No âmbito da mineração de texto, o terceiro item falou sobre os Sistemas de Recuperação de Informações, criado para sanar a necessidade de extração de informações de dados não estruturados. Além disso, o capítulo trouxe a definição do que são dados textuais não estruturados e também explicou de forma sucinta o processamento de linguagem natural.

O quarto item apresenta as etapas da mineração de texto, onde falou-se da coleta, pré- processamento, indexação, mineração e análise dos dados. acervo de

arquivos foi selecionado a fim de que seus dados fossem minerados. E, por fim, é seguido do sexto item,que mostra e analisa os resultados.

Em suma, entende-se que a mineração de textos é um método importante para que seja possível a extração de conhecimento de bases de dados não estruturadas, que são a forma de dados mais frequente nos dias de hoje.

REFERÊNCIAS

ANTONS, D. *et al.* The application of text mining methods in innovation research: current state,evolution patterns, and development priorities. **R&D Management**. 2020.

CIELEN, D; MEYSMAN, A; ALI, M. **Big data, machine learning, and more, using Python tools**. [s.l: s.n.]. 2016.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de Informação e Processamento da Linguagem Natural. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais** [...]. Porto Alegre: SBC, 2003. p. 347-395.

HASSANI, H. *et al*. Text mining in big data analytics. **Big Data and Cognitive Computing,** v.4, n.1, mar./2020. Disponível em: https://www.mdpi.com/2504-2289/4/1/1.

MORAIS, E.; AMBRÓSIO, A. **Mineração de textos**: relatório técnico. Goiania: INF/UFG, 2007.

PEZZINI, Anderson. Mineração de textos: conceito, processo e aplicações. **Revista Brasileira de Contabilidade e Gestão**, Ibirama, v. 5, n. 10, p. 58–61, 2023. DOI: 10.5965/2764747105102016058. Disponível em: https://revistas.udesc.br/index.php/reavi/article/view/6750. Acesso em: 27 fev. 2024.

PROVOST, F; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, v. 1, n.1, 13 fev. 2013. Disponível em: https://www.liebertpub.com/doi/10.1089/big.2013.1508

SERAPIÃO, P. R. B. *et al*. Uso de mineração de texto como ferramenta de avaliação dequalidade informacional em laudos eletrônicos de mamografia. **Radiol Bras**, v. 43, n. 2, abr./2010.