em questão
E-ISSN 1808-5245

Uso de deep learning para a construção de um modelo de recuperação da informação aplicado para o setor de mineração no Brasil

Luander Cipriano de Jesus Falcão [I]

[I] Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; luanderfalcao@yahoo.com.br; https://orcid.org/0000-0003-2417-6345

Brenner Lopes [II]

[II] Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; brenner.lopes@gmail.com; https://orcid.org/0000-0002-5807-0437

Renato Rocha Souza [III]

[III] Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; rsouzaufmg@gmail.com; https://orcid.org/0000-0002-1895-3905

Ricardo Rodrigues Barbosa [IV]

[IV] Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; rrbarb@gmail.com; https://orcid.org/0000-0003-3366-7525

Resumo: Diante do crescimento exponencial de dados e informações, proporcionado por sensores e mídias sociais, um ecossistema composto por novas infraestruturas de armazenamento e processamento, denominado *Big Data,* foi desenvolvido. Todo esse desenvolvimento redundou em uma nova área do conhecimento, denominada Ciência de Dados. Apesar de haver um ecossistema e uma área do conhecimento para tratar esse bloco massivo de dados e informação, o incomodo da superabundância de dados ainda permanece, e se torna mais expressivo quando as empresas tomam consciência que podem usar *zetabytes* de dados e informações para direcionarem a estratégia e as operações. Baseado nisso, essa pesquisa buscou desenvolver um método para resumir as notícias do setor de mineração do Brasil, identificando o efeito da similaridade semântica na análise, possibilitando a recuperação da informação e uso em processos de compreensão do setor. Nesse método foi aplicado o *transformer* BERTSUM para sumarizar as notícias, e após sumarizadas o *transformer* BERT foi aplicado para medir a similaridade entre as notícias. O método permitiu reduzir em 75% todo o bloco de texto, retirar notícias com o mesmo teor semântico, e deduzir que há um padrão no discurso das notícias relacionadas ao setor de mineração.

Palavras-chave: processamento de linguagem natural; deep learning; bert; ats; mineração

1 Introdução

Em um mundo de dados e informações, produzidos, armazenados, recuperados e compartilhados com maior velocidade a cada ano, a disponibilidade de acesso é iminente. Esta disponibilidade de dados e informações é gerada pelo crescimento exponencial da instalação de sensores, do poder computacional, das redes sociais, de Internet das Coisas (IoT) e de tecnologias móveis. Um exemplo dessa realidade é o aumento da taxa de geração de dados. Estima-se uma geração 2,5 quintilhões de dados, todos os dias, o que representa 250.000 vezes o tamanho da Biblioteca do Congresso Americano (Brands, 2014). Um estudo do International Data Corporation (IDC) de 2014, aponta que o universo digital dobra a cada dois anos e alcançou o tamanho de 40 *zetabytes* (ou 40 trilhões de *gigabytes*) em 2020, e esse universo digital, em 2013, era de 4,4 *zetabytes* (Nesi; Pantaleo; Sanesi, 2015). A esse fenômeno, da explosão de dados e informações, se dá o nome de *Big Data*.

Apesar de haver várias definições, há uma convergência de entendimento no sentido de ser "[...] especificamente um conjunto de dados tão grande ou complexo, que os aplicativos tradicionais de processamento de dados não são suficientes" (Jain; Gyanchandani; Khare, 2016, p. 1, tradução nossa) para registrar e processar tais dados. Para trabalhar com essas enormes e complexas bases de dados, tecnologias específicas foram desenvolvidas, principalmente para a arquitetura de dados (Balduini *et al*., 2019), que permitem o "[...] aumento da automação da coleta e análise de dados, junto com o desenvolvimento de algoritmos que podem extrair e ilustrar padrões em comportamentos humanos" (Aversa; Hernandez; Doherty, 2021, p. 2, tradução nossa).

Devido à capacidade de organização e recuperação da informação por meio das tecnologias atreladas ao *Big Data*, como a universalização da Inteligência Artificial (AI) (*Artificial Intelligence*), e as aplicações das técnicas analíticas no campo do *Machine Learning* (ML) e *Deep Learning* (DL), se originou um novo campo de estudo denominado Ciência de Dados.

O conceito de Ciência de Dados é dado como uma nova série de tecnologias, processos e sistemas para extrair valor e fazer descobertas (Wang,

2018), tendo como missão a transformação de dados brutos e confusos em conhecimento acionável (Stanton, 2012). Esse conceito traz uma visão mais pragmática para o *Big Data* por focar na coleta, preparação, análise, visualização, gerenciamento e preservação de grandes coleções de dados, não limitada apenas aos conhecimentos inerentes à ciência da computação, mas tendo interface com a matemática, estatística e outras áreas do conhecimento. Dada essas características da Ciência de Dados pode-se assumir ser esta interdisciplinar, possuindo maior capacidade de descobrir percepções, tendências valiosas e desconhecidas em um conjunto de dados de uma área do conhecimento (Salehi; Burgueño, 2018). Seu foco está na produção de *insights* analíticos e estabelecimento de modelos de previsão (Wang, 2018). Dentre as várias definições de Ciência de Dados, o entendimento proposto por Dhar (2013), merece destaque quando define essa abordagem como o estudo sistemático da organização, propriedades e análise de dados e seu papel na inferência.

Um dos pontos centrais de todo esse contexto envolve a recuperação da informação com foco no uso da informação. Este conceito, uso da informação, é de difícil definição (Choo, 2003), pois as necessidades informacionais não surgem plenamente formadas e podem atender a diversas finalidades. No caso de empresas, independente do ramo de negócios, a recuperação da informação e o seu uso tem por finalidade prover vantagem competitiva por meio da análise de dados, tanto para a área operacional quanto para a área estratégica da empresa (Rinaldi; Russo; Tommasino, 2021).

Quando as empresas tomam consciência de terem alta volumetria em seus bancos de dados, e esses podem ser conectados e disponíveis para gerar valor, surge o incômodo problema da sobrecarga de informações. O incômodo é potencializado quando se entende ser possível obter vantagem competitiva por meio de documentos textuais, ou por meio de informação textual em formato digital, que representa 80% da informação que circula na *Web* (Lamsiyah e*t al.*, 2021a). Porém, é um processo complexo, exaustivo e caro quando se utiliza recursos humanos, mesmo tendo um alto volume de informação disponível para

identificação e extração de ideias centrais e úteis dos textos (Mutlu; Sezer; Akcayol, 2020; Yang *et al*., 2013).

Sendo essa a direção estratégica proporcionada pelas tecnologias de *Big Data* e pela Ciência de Dados, as empresas passaram a focar nos novos desafios relacionados à organização dos dados. Esses desafios incluem, principalmente, a necessidade de examinar os fluxos informacionais e as aspirações do negócio que orientam o uso das informações (Weaver, 2021). Mas a resolução desses desafios esbarra na carência de mão de obra qualificada, para trabalhar no ecossistema do *Big Data*, falta de cientistas de dados, para aplicar os conceitos de Ciência de Dados e a falta de tempo necessário, para interpretar os dados resultantes dessas tecnologias. Aliado a isso ainda há a incapacidade das pessoas em assimilarem inúmeras informações, inclusive na forma de texto não estruturado, tornando métodos de resumo de texto eficientes e importantes (Padmakumar; Saran, 2016).

Partindo da premissa de que em "[...] todos os segmentos de negócio, uma empresa chegará à liderança através do uso da informação como uma arma competitiva, e no processo, mudando as regras da competição para todo mundo" (McGee; Prusak, 1994, p. 71), se explica o uso intensivo de Ciência de Dados.

O uso intensivo de Ciência de Dados no setor de mineração no Brasil é justificado pela importância deste em termos de participação no PIB, geração de empregos, produção e pela ramificação no Brasil por meio dos elos da cadeia produtiva. Além disso há o fato de haver grandes repositórios textuais com informações antigas e não mais utilizadas com potencial de serem transformadas em informações preciosas para a organização (Ramos; Bräscher, 2009). Todo esse conjunto de fatores justifica a aplicação das modernas ferramentas de Ciência de Dados utilizadas para tratamento de texto.

Dada a importância do setor e consequentemente a vasta produção de material textual em torno dele, a mineração é um setor muito dinâmico. Dinâmico no sentido de haver "[...] alta intensidade de informação que pode indicar mudança considerável" (Miller, 2002, p. 43). Ao assumir que notícias são informações, e quanto maior a quantidade de notícias maior será a intensidade informacional de um setor, e consequentemente, mais dinâmico.

Essas notícias podem ser verdadeiras ou falsas, positivas ou negativas, podem cumprir um objetivo estratégico ou simplesmente informar algo de interesse geral. O conjunto dessas notícias é capaz de mudar o entendimento, a compreensão do usuário sobre uma temática específica do setor, ou focar a sua necessidade informacional em um alvo. Essas notícias podem modificar ou criar sentido sobre o setor e em seus respectivos aspectos.

Diante dessa complexidade surge a necessidade de construção de modelos capazes de recuperar informações mais aderentes ao uso final e de apresentá-las de forma eficiente e eficaz.

Nesse contexto, o objetivo de pesquisa deste trabalho é tratar o excesso de informação textual disponível de forma a gerar um novo patamar de compreensão do setor de mineração no Brasil.

Exposto isso a pergunta central que esse trabalho pretende responder é: Como recuperar e analisar informações do setor de mineração no Brasil a partir de notícias?

Dente os resultados obtidos pode-se citar a capacidade da sumarização de texto ao permitir o uso de outras técnicas, a identificação da repetição semântica mascarando termos importantes, e a influência da data da notícia, indicando um efeito longitudinal na análise.

2 Referencial teórico

A análise de texto e linguagem por meio computacional é dado o nome de Processamento de Linguagem Natural (PNL), ou *Natural Language Processing* (NLP), e sua base está na interdisciplinaridade de conceitos encontrados na ciência da computação, ciência da informação, lingüística, matemática, inteligência artificial, lógica e psicologia (Chowdhury, 2003; Joshi, 1991).

O NLP possui diversos modelos de desempenho em uma variedade de tarefas, como análise de sentimentos, tradução de idiomas, reconhecimento de entidades de nomes e resumo automático de texto. O Resumo Automático de Texto possui outras denominações, como Sumarização de Texto, *Automatic Summarization*, Sumarização Automática de Texto, e *Automatic Text*

Summarisation (ATS), sendo esse o termo mais utilizado em NLP. As abordagens relativas ao conceito de ATS permitem várias aplicações, como:

Quadro 1 - Principais características da sumarização de texto

(1) Redução do tempo de leitura.
(2) Facilitação do processo de seleção e pesquisa de documentos.
(3) Melhora na eficácia da indexação.
(4) Redução de vieses encontrados em sumarizadores humanos.
(5) Aplicação em sistemas de perguntas/respostas.
(6) Maior capacidade de processamento de documentos.

Fonte: Elaborado pelos autores.

O ATS visa reduzir grandes blocos de texto, em textos menores, na forma de resumos abrangentes e concisos, capazes de reter as informações mais relevantes, críticas e úteis para obter uma melhor compreensão do texto original, sem perder o sentido original do mesmo (Goularte *et al.*, 2014, 2019; Syed; Gaol; Matsuo, 2021; Tan; Kieuvongngam; Niu, 2020; Yang *et al*., 2013). Com o ATS é possível resumir um conteúdo longo de notícias em uma versão mais curta (Protim Ghosh; Shahariar; Hossain Khan, 2018).

2.1 Caracterização do resumo automático de texto

O termo Sumarização de Texto (*Text summarization*) precede o termo Resumo Automático de Texto, ou Sumarização Automática de Texto (*Automatic Text Summarization* em inglês). O *Text Summarization* (TS) é o processo de condensar o texto fonte em uma versão mais curta, preservando seu conteúdo de informação e significado geral (Padmakumar; Saran, 2016). Ao reduzir o conteúdo do texto, o *Text Summarization* pode impactar negativamente o significado transmitido deste (Yang *et al*., 2013), por isso as técnicas utilizadas normalmente empregam vários mecanismos para identificar sentenças altamente relevantes no texto ou remover frases/sentenças redundantes (Padmakumar; Saran, 2016).

Para Joshi *et al.* (2019, p. 200, tradução nossa) *Automatic text summarization* "[...] visa representar documentos de textos longos de forma

compactada para que as informações possam ser rapidamente compreendidas e legíveis para os usuários finais". De acordo com Alami, Meknassi e En-Nahnahi (2019, p. 195, tradução nossa), o "[...] objetivo é a produção de uma versão abreviada de um grande documento de texto, preservando a ideia principal existente no documento original". A definição de ATS é praticamente a mesma de TS, porém, ATS foca em documentos de texto maiores, enquanto TS foca mais no método, na redução do texto.

Existem várias técnicas de ATS e a maioria dessas abordagens modelam o resumo de texto como um problema de classificação que resulta em incluir ou não a frase no resumo, por métodos de pontuação da frase baseado em recursos estatísticos como TF-IDF (Hark; Karci, 2020; Sinha; Yadav; Gahlot, 2018) e outras funções de pontuação. Um exemplo mais complexo dessa pontuação é encontrado em John, Premjith e Wilscy (2017) quando propuseram um sistema de sumarização multi-documento não supervisionado extrativo a partir de sentenças salientes dos documentos de entrada, considerando a sumarização como um problema de otimização multicritério.

Para Christian, Agus e Suhartono (2016) há três aspectos importantes que caracterizam a pesquisa sobre ATS: (1) o resumo pode ser produzido a partir de um único documento ou de vários documentos; (2) o resumo deve preservar informações importantes; e (3) o resumo deve ser curto. Padmakumar e Saran (2016) classificam o ATS quanto a extensão do resumo, que pode ser para criar um título ou um conjunto de palavras-chave e para gerar uma sequência curta, mas coerente de frases.

O ATS pode ser classificado em tipos diferentes. Quando baseado no tipo de saída do resumo pode ser classificado como *Extractive* ou *Abstractive*. Quando baseado na quantidade de documentos de entrada pode se classificado como *Single Document* ou *Multi Document*. Quando baseado no propósito pode ser do tipo *Generic, Query-based e Domain-specific*.

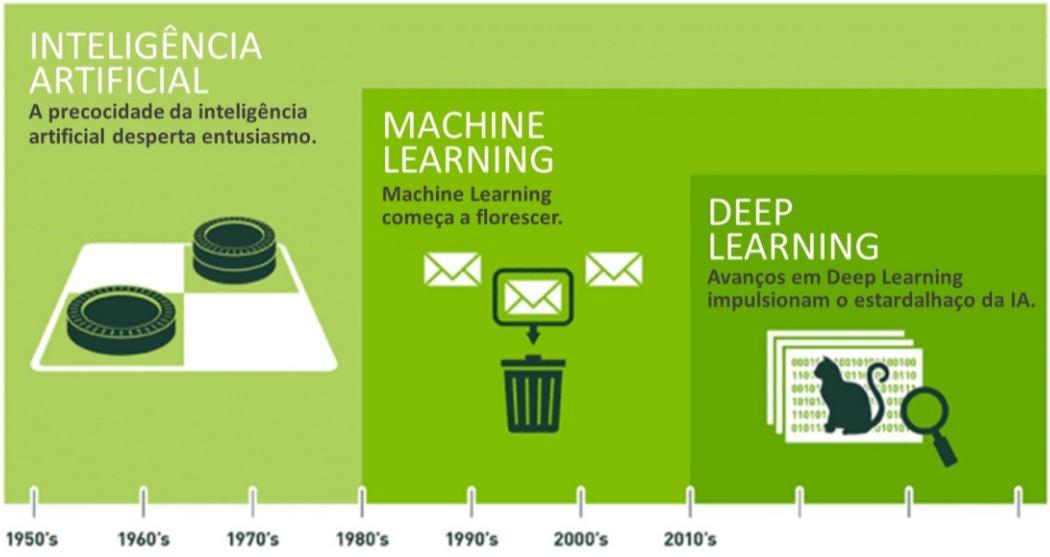
2.2 Uso de deep learning para resumo automático de texto

A Inteligência Artificial oferece vantagens para lidar com problemas complexos associados a incertezas, agilizar o processo de tomada de decisão, diminuir as

taxas de erro e aumentar a eficiência computacional (Salehi; Burgueño, 2018). Entre as diferentes técnicas de Inteligência Artificial estão o *Machine Learning* (ML) e o *Deep Learning* (DL). Entretanto muitas vezes o termo Inteligência Artificial é confundido com *Machine Learning* e *Deep Learning*, ou *Machine Learning* é usado como sinônimo para *Deep Learning* (Hamet; Tremblay, 2017).

Uma forma mais útil de pensar sobre o relacionamento dos termos é visualizá-los como círculos concêntricos com Inteligência Artificial, sendo esse o maior, por ser a primeira ideia, e então *Machine Learning*, por ter florescido depois, e finalmente *Deep Learning* contendo parte das duas (Copeland, 2016). Essa visão está representada na Figura 1 abaixo:

Figura 1 - A Correlação entre Inteligência Artificial, *Machine Learning* e *Deep Learning*



Fonte: Copeland (2016).

O *Deep Learning* (DL) é um ramo, um tipo específico do *Machine Learning*, consequentemente da Inteligência Artificial, que tende a aprender as múltiplas representações de dados (Goodfellow; Bengio; Courville, 2016; Khamparia; Singh, 2019; Leijnen; Veen, 2020; Salehi; Burgueño, 2018). O *Deep Learning* (DL) permitiu muitas aplicações práticas de *Machine Learning* e consequentemente de Inteligência Artificial, pois o seu impacto foi percebido em quase todos os campos científicos, e tem transformado negócios e indústrias (Shrestha; Mahmood, 2019). Em termos de compreensão da evolução histórica

do *Deep Learning*, Goodfellow, Bengio e Courville (2016) identificaram as principais características, conforme Quadro 2 abaixo:

Quadro 2 - Principais características da evolução histórica do *Deep Learning*

O *Deep learning* tem uma longa e rica história, tem muitos nomes refletindo diferentes pontos de vista filosóficos, e aumentou e diminuiu em popularidade.
O *Deep learning* se tornou mais útil conforme aumentou a disponibilidade de modelos treinados.
O crescimento dos modelos de *Deep learning* aumentaram ao longo do tempo devido a melhoria da infraestrutura dos computadores, tanto de *hardware* quanto de *software*.
O *Deep learning* resolveu problemas de aplicações cada vez mais complicadas com o aumento da precisão ao longo do tempo.

Fonte: Adaptado de Goodfellow, Bengio e Courville (2016, tradução nossa).

Uma arquitetura de *Deep Learning* (DL) é baseada em redes neurais profundas (*Deep Neural Networks*), ou seja, redes neurais com mais de uma camada oculta, nas quais o aumento do número de camadas resulta em uma rede mais profunda (Salehi; Burgueño, 2018). A modelagem de *Deep Neural Networks* (DNN) é baseada em um número maior de camadas de processamento, permitindo que funções mais complexas e não lineares sejam mapeadas, dando origem as modernas arquiteturas de *Deep Learning*, também denominadas como redes neurais convolucionais (*Convolutional Neural Networks* - CNNs), redes neurais recorrentes (*recurrent neural networks* - RNNs), *autoencoders*, redes de crenças profundas (*deep belief nets*) e outras (Salehi; Burgueño, 2018; Shrestha; Mahmood, 2019).

Existem várias arquiteturas de *Deep Learning* para a representação de texto relacionadas a diferentes tarefas de NLP, como Word2Vec (*Word Embedding*), CBOW (*The Continuous Bag of Words Model*), GloVe (*Word Embedding with Global Vectors*), FastText (*Subword Embedding*), BERT (*Bidirectional Encoder Representations from Transformers*), CoVe (*Context Vectors*), ELMo (*Embeddings from Language Models*), GRU (*Gated Recurrent Units*), entre outros (Goodfellow; Bengio; Courville, 2016; Zhang e*t al.*, 2023).

Essas arquiteturas também são chamadas de *Transformers.* A arquitetura do *Transformer* é dimensionada com dados de treinamento e tamanho do

modelo, facilitando o treinamento paralelo eficiente e capturando recursos de sequência de longo alcance, além de serem bibliotecas *open-source* (Wolf *et al*., 2019). Os *transformers* são um avanço para tarefas de aprendizagem de sequência, introduzidas pelo Google em 2017, e são baseados inteiramente em mecanismos de atenção, eliminando assim a necessidade de unidades recorrentes e também de convolução, por possuírem uma arquitetura com um codificador e um decodificador que são empilhados várias vezes (Syed; Gaol; Matsuo, 2021).

2.2.1 O *Transformer* BERT e sua derivação BERTSUM

O modelo *Bidirectional Encoder Representations from Transformers* (BERT) é baseado em *um multi-layer bidirectional transformer* com *attention mechanisms*, desenvolvido por Devlin *et al.* (2018). Apesar de ser relativamente recente consta em vários trabalhos de pesquisa, devido a sua versatilidade, tanto de processamento de muitos dados, quanto por permitir multi-idiomas e multi-tarefas de NLP.

O BERT é treinado em uma grande quantidade de dados de origem (cerca de 3300 milhões de palavras) da *Wikipedia,* no idioma inglês, e *BookCorpus*, usando duas tarefas não supervisionadas, incluindo a *masked language modelling* e a previsão da próxima frase (Lamsiyah *et al.*, 2021b). O modelo pré-treinado pode ser aplicado a uma nova tarefa de processamento de linguagem natural, adicionando algumas camadas ao modelo de origem, como classificação de texto, sistemas de perguntas/respostas e sumarização automática de texto.

Além do seu desempenho superior a outros algoritmos de NLP na incorporação de sentenças, a arquitetura BERT foi selecionada por se basear na arquitetura do *transformer* e ter sido desenvolvida com objetivos específicos para o pré-treinamento. Em uma etapa, ele mascara aleatoriamente 10% a 15% das palavras dos dados de treinamento, tentando prever as palavras mascaradas, e a outra etapa recebe uma frase de entrada e uma frase candidata, prevendo se a frase candidata segue corretamente a frase de entrada (Devlin *et al*., 2018; Miller, 2019). Esse processo pode levar vários dias para treinar, mesmo com

uma quantidade substancial de GPUs. Devido a este fato, o Google lançou dois modelos BERT para consumo público, onde um tinha 110 milhões de parâmetros e o outro continha 340 milhões de parâmetros (Devlin *et al*., 2018; Miller, 2019).

Liu e Lapata (2019) adotaram BERT para sumarização de texto, porém ajustaram o *transformer,* uma vez que, para sumarização, a sua aplicação não é direta. O BERT é treinado como um modelo de linguagem mascarada, nos quais os vetores de saída são fundamentados em *tokens* em vez de sentenças, enquanto na sumarização extrativa, a maioria dos modelos manipula representações em nível de sentença. Embora os *embeddings* de segmentação representem sentenças diferentes, no BERT, eles só se aplicam a entradas de pares de sentenças, enquanto na sumarização deve-se codificar e manipular entradas multissentenciais. A esse modelo os autores chamaram de BERTSUM.

2.3 Revisão do estado da arte em resumo automático de texto

Segundo Vasconcellos, Silva e Souza (2020, p. 2), "[...] o Estado da Arte e o Estado do Conhecimento são denominações de levantamentos sistemáticos ou balanço sobre algum conhecimento, produzido durante um determinado período e área de abrangência". Diante desse princípio a estratégia de busca concentrou-se no que havia de mais relevante em termos de Sumarização de Texto e *Deep Learning* por meio do *Transformer* BERT.

Abdel-Salam e Rafea (2022) desenvolveram um estudo sobre o desempenho de variantes de modelos baseados em BERT na sumarização de texto, por meio de uma série de experimentos, e propuseram o "SqueezeBERTSum". Esse modelo de sumarização, treinado e ajustado com a variante codificadora SqueezeBERT, alcançou pontuações competitivas do ROUGE mantendo o desempenho do modelo de linha de base BERTSUM em 98%, com 49% menos parâmetros treináveis. Dado os resultados dos experimentos na metodologia, existe uma versão potencializada do resumidor extrativo SqueezeBERT a partir dos resultados registrados acima.

Bondielli e Marcelloni (2021) propuseram uma metodologia para representar o perfil de currículos profissionais de candidatos a emprego, baseada

em arquiteturas de sumarização e *transformers* para geração de *embeddings* de currículos e em algoritmos de agrupamento hierárquico para agrupar esses *embeddings*. Optaram por trabalhar com BERT (*Bidirectional Encoder Representations from Transformers*) por esse tipo de *transformer* ser capaz de obter melhores resultados em uma ampla gama de tarefas de NLP.

Searle *et al*. (2021) fizeram um exame quantitativo da redundância de informações em notas EHR (*Electronic Health Records*) para avaliarem inovações que operam em narrativas clínicas. Primeiro estimaram a entropia da linguagem clínica usando GPT-2, que é um modelo anterior de linguagem causal auto-regressivo de última geração, baseado na arquitetura de *Transformer*. Depois utilizaram um segundo método para estimar os níveis de redundância em um texto clínico, aplicando métricas de avaliação de sumarização a pares de notas ordenados, por meio de BERT.

Lamsiyah *et al*. (2021b) ao estudarem *Transfer Learning* (Transferência de Aprendizagem) usando modelos pré-treinados de *word embedding* em *text summarization* perceberam que a maioria das representações não considera a ordem e as relações semânticas entre as palavras em uma frase e, portanto, não carregam o significado de uma frase inteira. Para contornar esse problema propuseram um método não supervisionado para a sumarização extrativa de múltiplos documentos com base em *Transfer Learning* (Transferência de Aprendizagem) a partir do modelo de *embedding* de frases de BERT.

Li e Yu (2021) apresentaram um modelo de sumarização extrativa baseado em BERT e em uma rede de memória dinâmica (*dynamic memory network*). Os autores utilizaram o *transformer* do BERT para extrair recursos de texto e para construir os *embeddings* de frases a partir do modelo pré-treinado. O modelo baseado em BERT rotula as frases automaticamente sem usar nenhum recurso artesanal e os conjuntos de dados são rotulados de forma simétrica. Resultados experimentais mostraram que o modelo, baseado em BERT e rede de memória dinâmica, alcança o resultado comparável com outros sistemas extrativos nos conjuntos de dados.

A busca na literatura acadêmica resultou em poucos trabalhos que refletem o tema dessa pesquisa. As pesquisas apresentadas acima mostram

intensivo uso de BERT para sumarização de texto, porém, a maioria se refere a criação de uma metodologia de sumarização ou ao melhoramento do BERT, para se tornar mais performático na tarefa de sumarização. Poucas pesquisas podem ser consideradas como aplicadas, ao exemplo de Bondielli e Marcelloni (2021) e Searle *et al*. (2021), que utilizaram as metodologias de sumarização em BERT para identificarem um fator ou uma característica dentro de um bloco textual. Não foi identificado nenhum trabalho que trate especificamente sobre a elaboração de uma metodologia de recuperação da informação, a partir da aplicação de sumarização automática de texto focado em um setor.

3 Metodologia

Esse trabalho foi motivado pela variedade de técnicas de Sumarização Automática de Texto, as quais permitem explorar um conjunto de dados textuais em um contexto prático, direcionando a busca e o uso da informação de uma forma mais racional.

Em termos metodológicos essa pesquisa se classifica como indutiva, por partir "[...] de dados ou observações particulares constatadas, podendo chegar a proposições gerais" (Richardson, 2012, p. 35). Quanto à finalidade/natureza como aplicada, pois o objetivo é "[...] gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos" (Prodanov; Freitas, 2013, p. 51). E em relação aos objetivos como exploratória, por visar o exame de um "[...] conjunto de fenômenos, buscando anomalias que não sejam ainda conhecidas e que possam ser, então, a base para uma pesquisa mais elaborada" (Wazlawick, 2014, p. 22). Quanto à abordagem do problema, a pesquisa se classifica como método misto ou quali-quantitativo, por usar concomitantemente, métodos quantitativos e qualitativos (Creswell; Creswell 2014).

A classificação da pesquisa permite direcionar esforço para enfrentar o desafio de lidar com massivas quantidades de dados textuais, em grandes conjuntos de dados digitais. Diante do exposto esse trabalho de pesquisa tem como ponto de partida se apropriar das modernas técnicas de NLP que utilizam *Deep Learning*. O desenvolvimento deste trabalho foi guiado pelo uso da tarefa

de NLP denominada ATS (*Automatic Text Summarisation*) aplicada à notícias referentes ao setor de Mineração no Brasil.

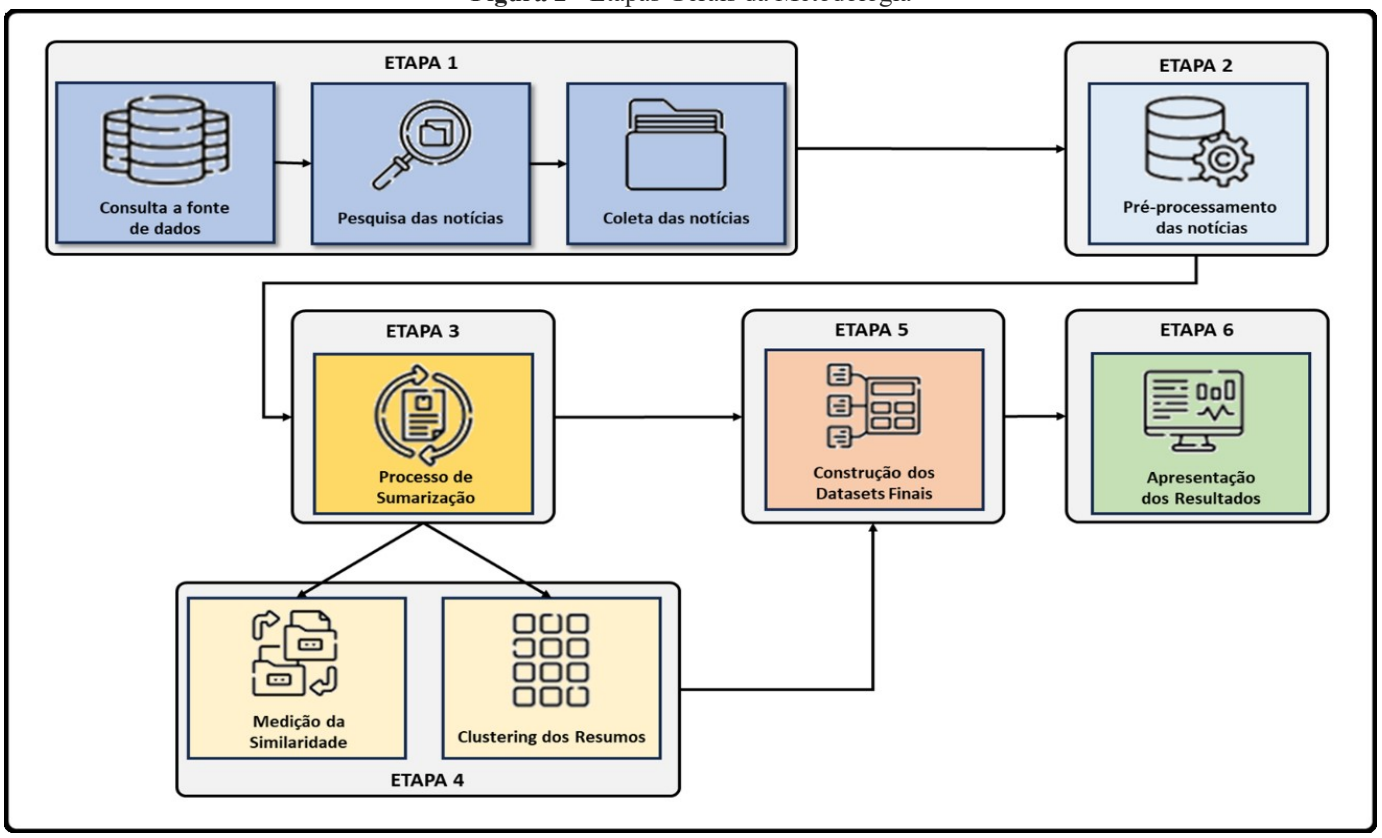
Enquanto escopo de pesquisa esse trabalho abordará o uso de notícias no idioma português e inglês, e não adotará o uso de um fluxo dinâmico de entrada e tratamento de notícias. Será tratado um bloco estático de notícias.

3.1 Etapas da metodologia

Este trabalho apresenta um método de recuperação da informação a partir da aplicação de sumarização automática de texto do setor de mineração no Brasil.

Para alcançar este objetivo foi utilizado o *Transformer* BERTSUM para a tarefa de Sumarização de Texto, BERT para mensurar o Grau de Similaridade Textual, e a técnica de *Machine Learning* de Clusterização por meio de *Bag of Word* (BOW).

Figura 2 - Etapas Gerais da Metodologia



Fonte: Elaborado pelos autores.

A Figura 2 fornece uma visão geral da metodologia e das 6 etapas que a compõem, sendo: (1) Pesquisa e Coleta das Notícias; (2) Pré-processamento das

Notícias; (3) Processo de Sumarização; (4) Tarefas Secundárias de NLP; (5) Construção dos *Datasets* para Análise; e (6) Apresentação dos Resultados.

As duas primeiras etapas são responsáveis pela preparação dos dados, a segunda pela sumarização das notícias, enquanto as demais pela preparação das bases de dados para análise dos resultados. A seguir, cada uma das etapas é descrita em detalhes.

A primeira etapa consiste em extrair um *corpus* de dados, sendo cada notícia um documento, para a realização do processamento estabelecido. As notícias alvo do setor de mineração do Brasil foram obtidas de uma base de uma empresa proprietária de notícias. Essa empresa proprietária monitora mais de dez milhões de perfis de empresas, 197 mercados, mais de 370 setores, possui mais de 450.000 relatórios de pesquisa, incorpora mais de 40.000 novas histórias por dia, possui mais de 2.000 fontes de informações indexadas e 16 idiomas. Devido a abrangência de países, setores e fontes essa empresa detentora dessa fonte de dados foi escolhida. A *query* de consulta utilizada na plataforma da empresa está descrita na Fórmula 1 (Quadro 3) abaixo e as notícias retornadas dessa pesquisa serão exportadas, cada uma para um arquivo de texto.

Quadro 3 - *Query* de consulta

(mineração or mining) not (criptomoedas or cripto or bitcoin or "data mining" or "mineração de dados" or "process mining" or "mineração de processos")

Fonte: Elaborado pelos autores.

A segunda etapa, Pré-processamento das Notícias, cada arquivo de notícia foi inserido em um único arquivo de texto, para adequação da estrutura, de forma que todas as reportagens seguissem o mesmo padrão, sendo na primeira linha a data da notícia, na segunda linha a fonte da notícia, na terceira linha o título da notícia e na quarta linha a notícia na integra. Esse processo permitiu retirar notícias que tinham apenas data e título. O passo seguinte foi transformar cada notícia em um arquivo em pdf, e construir um fluxo de importação de cada notícia em pdf para um arquivo tabular. Nesse arquivo

tabular foram imputados a data da notícia, a fonte, o título e o tamanho da reportagem em quantidade de caracteres.

Na etapa três a linguagem de programação escolhida foi o *Python*, e para processar os resumos foi escolhido o BERTSUM, uma variação do BERT desenvolvido especificamente para sumarização extrativa. A notícia do arquivo pdf foi extraída e na sequência sumarizada. Para cada referência de arquivo pdf, na tabela de dados, foi imputada uma sumarização da notícia original, e um novo campo contendo o tamanho da sumarização em termos de quantidade de caracteres.

Na etapa quatro, Tarefas Secundárias de NLP, foram aplicados dois processos de NLP, sendo: (a) Medição da Similaridade Textual para cada resumo; e (b) *Clustering* dos Resumos.

A Similaridade Textual é utilizada para determinar o quão semelhante dois textos são. A saída é uma pontuação de similaridade entre zero e um, sendo que, quanto mais próximo de um mais semelhantes os textos são. Para essa tarefa foi construído um fluxo no qual a primeira notícia da tabela foi comparada com as demais notícias, gerando uma pontuação de similaridade para cada par de notícias. A pontuação de similaridade de cada par de notícias foi classificada, observado a seguinte regra descrita no Quadro 4:

Quadro 4 - Classificação da Pontuação de Similaridade

Pontuação de Similaridade	Tipo de Similaridade
0,950 a 1,000	Perfeita
0,750 a 0,949	Forte
0,500 a 0,749	Moderada
0,250 a 0,499	Fraca
0,100 a 0,249	Ínfima
0,00 a 0,099	Nula

Fonte: Elaborado pelos autores.

A partir dessa classificação dos pares de notícias foi contado a quantidade de notícias por tipo de similaridade por notícias. Exemplo do resultado da contagem: a notícia de código 3809 teve cinco pares classificados

como Tipo de Similaridade Nula, 35 Ínfima, 1.229 Fraca, 1.879 Moderada, 75 Forte e 0 Perfeita. Quando a contagem dos pares de notícias for igual a zero na classificação Perfeita significa que essa notícia não é similar a nenhuma outra com uma pontuação ente 0,95 e 1,00.

O último processo aplicado foi o de *Clustering*, visando agrupar as notícias por semelhança de discurso, com o objetivo de captar a história contada por um grupo de notícias. Para o *Clustering* foi escolhida a técnica de *K-means*. Em algoritmos de *Cluster* a métrica *Silhouette* indica o quanto os *clusters* estão separados ou não. O valor da pontuação *Silhouette* varia de -1 a 1. Se a pontuação for um, o *cluster* é denso e bem separado dos outros *clusters*. Um valor próximo a 0 representa *clusters* sobrepostos com amostras muito próximas ao limite de decisão dos *clusters* vizinhos. Uma pontuação negativa indica que as amostras podem ter sido atribuídas aos *clusters* errados.

Na quinta etapa, Construção dos *Datasets* para Análise, foram obtidos dois arquivos de dados, sendo um composto por uma chave única para cada notícia, a sumarização dessa notícia, e quantas notícias por tipo de similaridade. O segundo arquivo também possui uma chave única para cada notícia, de forma que os arquivos possam ser relacionados, além de um campo para indicar o contexto, o *cluster* e as dez palavras mais relevantes por cluster. E a sexta etapa é a Apresentação dos Resultados.

4 Discussão e resultados da pesquisa

Ao todo foram coletadas 3.824 notícias datadas entre 11/02/2003 e 30/11/2021, e inseridas em uma tabela conforme descrito na metodologia. Foram retiradas as notícias com 0 caracteres e aquelas que possuíam o título idêntico a outras, ficando ao final 3.271 notícias. Ao agregar a quantidade de notícias por ano, optou-se por retirar as notícias entre os anos de 2003 e 2012, devido a baixa quantidade de notícias nestes anos. Ao fim dessa etapa se obteve 3.224 notícias.

As 3.224 notícias selecionadas totalizaram 8,6 milhões de caracteres sendo a variação de caracteres por notícia entre 77 e 49 mil caracteres, demonstrando a alta variabilidade do tamanho das notícias captadas. Ao aplicar o sumarizador houve uma redução de 75% do texto original, de 8,6 milhões de

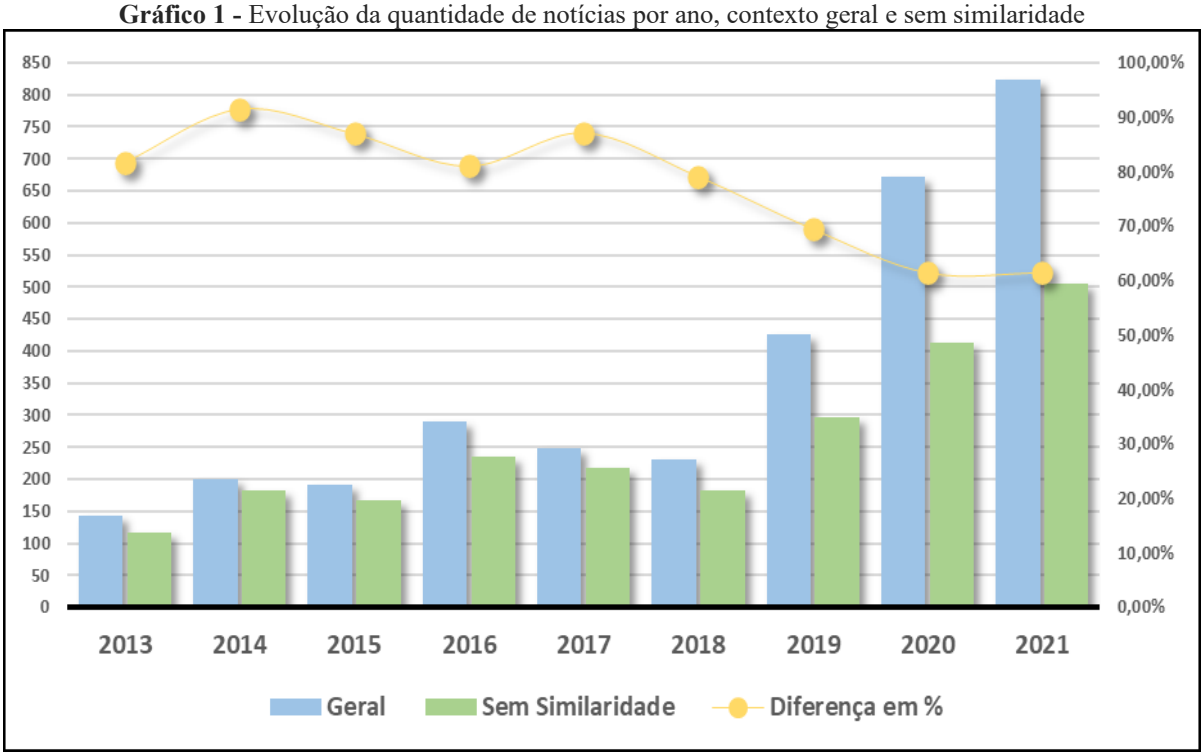
caracteres para 2,2 milhões de caracteres, com as notícias tendo redução entre 9% e 99%. Esse fato corrobora o ATS ser uma etapa anterior e importante no processamento textual, por reduzir o texto e possibilitar a utilização de outras tarefas.

Cada mensuração de similaridade foi inserida no arquivo com o seu respectivo par de notícia. Foi considerado que quando uma notícia tiver Tipo de Similaridade Perfeita igual a 0 (zero), significa que ela não possui similaridade com nenhuma outra notícia. Ao retirar as notícias com alto teor de similaridade se obtêm 2.315 notícias, indicando que nessas notícias não há repetição semântica, logo, não sendo necessário acessar todos os documentos para recuperar a informação. Nesse ponto há a corroboração de uma das vantagens do ATS, a redução da quantidade de documentos acessados para recuperar a informação.

Ao analisar a evolução da quantidade de notícias por ano, tanto no contexto Geral, que corresponde à 3.224 notícias, quanto no contexto Sem Similaridade, que corresponde à 2.315 notícias, os dados mostram um crescimento da quantidade de notícias entre os anos de 2013 e 2016, seguido de um declínio na quantidade de notícias em 2017 e 2018, justificado em parte pela retração da economia proveniente dos anos de 2015 a 2017.

A quantidade das notícias volta a crescer em 2019 devido, em parte, pela retomada da economia, e com reflexo no setor de Mineração. Os dados mostram que a partir de 2018 a quantidade de notícias com o mesmo teor semântico passa a aumentar. Isso é constatado dividindo a quantidade notícias do contexto Sem Similaridade pela quantidade de notícias do contexto Geral por ano. Essa proporção reduz a partir de 2018, indicando uma maior repetição semântica das notícias, pois as fontes de notícias podem publicar a mesma notícia e na mesma data. Esses resultados podem ser observados no Gráfico 1 abaixo:

Gráfico 1 - Evolução da quantidade de notícias por ano, contexto geral e sem similaridade



Fonte: Elaborado pelos autores.

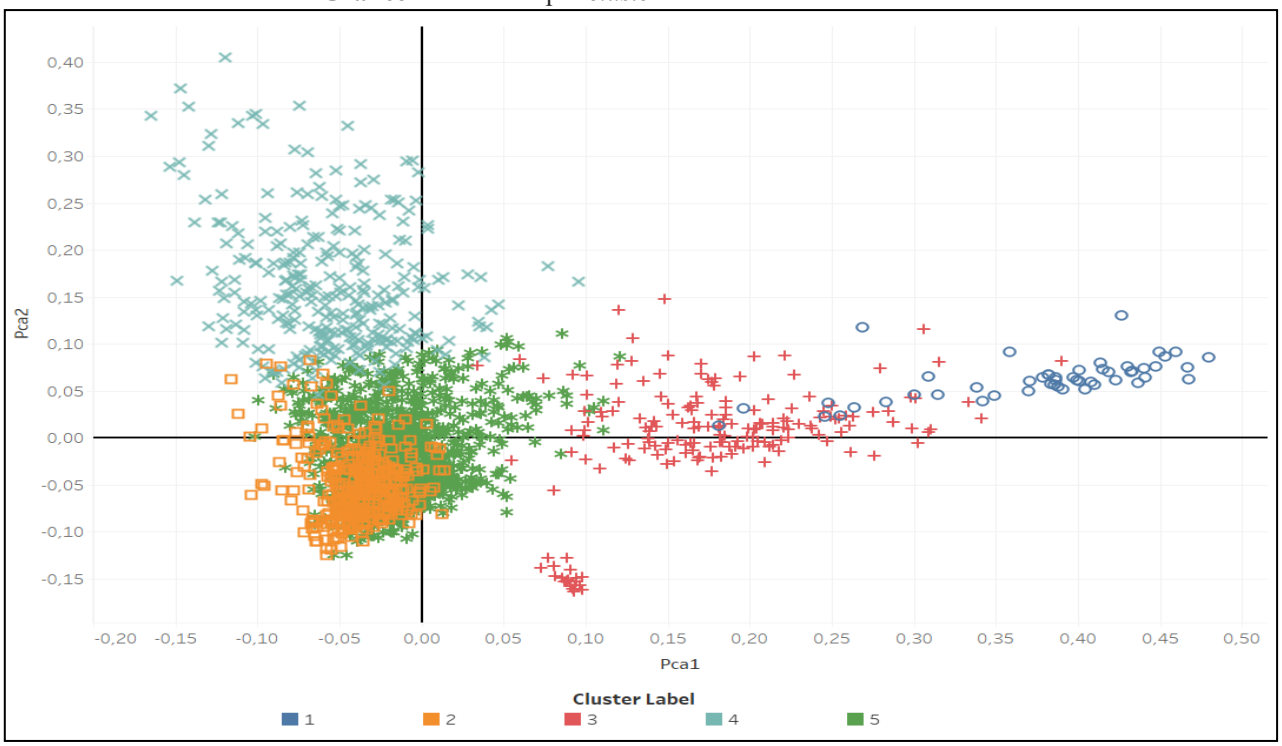
O aumento da repetição das notícias por meio de outras fontes de notícias no mesmo período fica mais evidente quando se analisa a quantidade de notícias por mês, no contexto Geral e no contexto Sem Similaridade. Esse fato endossa a dedução de uma maior repetição semântica do conteúdo das notícias a partir do ano de 2019. A evolução da quantidade de notícias mostra uma influência do tempo, ou fatores ligados ao período que devem ser considerados na análise.

4.1 Apresentação e análise dos resultados no contexto sem similaridade

No contexto Sem Similaridade há 2.315 notícias e foram gerados cinco *Clusters*. Foram criados dois eixos por meio da Análise de Componentes, conforme gráfico 2. A pontuação *Silhouette* foi de 0,0052 evidenciando uma forte sobreposição entre os *Clusters*. Essa sobreposição ocorre por causa das notícias do *Clusters* dois e cinco, por serem os mais representativos. Juntos representam cerca de 77% das notícias. As notícias desses *Clusters* estão mais próximas umas das outras, enquanto os demais *Clusters* possuem uma dispersão maior, compartilhando algumas características com os demais. O *Cluster* 3 compartilha

características com o *Cluster* cinco e com o um, e possui um bloco a parte. O resultado do gráfico 2 juntamente com a pontuação *Silhouette* indica uma padronização no discurso dos *Clusters*, mesmo quando a similaridade semântica é retirada.

Gráfico 2 - Notícias por *cluster* Contexto Sem Similaridade



Fonte: Elaborado pelos autores.

Cada *Cluster* é formado por uma quantidade de notícias e caracterizado pelas dez palavras mais representativas, conforme descrito no Quadro 5 a seguir:

Quadro 5 - Principais palavras do Contexto Sem Similaridade

Termos	**Clusters**				
	1	**2**	**3**	**4**	**5**
Termo 01	Tóquio	Mineração	Ibovespa	Minério	Mineração
Termo 02	bolsa	barragem	pontos	ferro	empresa
Termo 03	registradas	minas	índice	ano	*mining*
Termo 04	fechou	vale	*brazilian*	trimestre	brasil
Termo 05	valores	barragens	alta	milhões	projeto
Termo 06	paulo	nacional	baixa	produção	companhia

Termo 07	maiores	agência	ganhos	toneladas	setor
Termo 08	ações	estado	ações	bilhões	mineradora
Termo 09	altas	anm	queda	ouro	vale
Termo 10	sessão	região	*machinery*	vale	mercado

Fonte: Elaborado pelos autores.

O *Cluster* 1 possui cerca de 51 notícias, ou 2,20% do total. Ao analisar as palavras mais representativas desse *Cluster*, as notícias tratam basicamente da bolsa de valores de Tóquio, e de empresas ligadas a mineração do Japão. Dada a repetição dos termos se tornou um *Cluster* isolado. Foi denominado *Bolsa de Tóquio*.

O *Cluster* 2 possui 352 notícias, cerca de 15,21% do total, sendo o segundo maior *Cluster* em quantidade de notícias, e é caracterizado pela temática Agência Nacional de Mineração (ANM), barragem e estado de Minas Gerais. Cerca de 84% das notícias estão concentradas nos anos de 2019, 2020 e 2021. Esse *Cluster* trata do rompimento da barragem da Vale no município de Brumadinho, no estado de Minas Gerais em janeiro de 2019, as ações e acompanhamentos tomados após o rompimento da barragem, e a ameaça que outras barragens representavam. Dado o conjunto de palavras desse *Cluster* e do *Cluster* 5 se justifica em parte o fato de ambos os *Clusters* terem forte sobreposição. Foi denominado de *Barragem*.

O *Cluster* 3 possui cerca de 180 notícias, ou 7,78% do total. Ao analisar as palavras mais representativas desse *Cluster*, as notícias tratam do comportamento da Bolsa de Valores de São Paulo, variações tanto de alta quanto de baixa, ganhos e perdas. Nesse *Cluster* as notícias dos anos de 2013 e 2014 representam 63%, mascarando outros assuntos e evidenciando o efeito longitudinal na análise. O termo *machinery* representa a parte do *Cluster* mais afastada dos eixos, por tratar de notícias no idioma inglês. Foi denominado de *Ibovespa*.

O *Cluster* 4 possui cerca de 291 notícias, ou 12,57% do total. Ao analisar as palavras mais representativas, as notícias tratam da produção de minério de ferro, por ano e por trimestre, na quantidade de milhões e bilhões, pela empresa Vale. Também aborda a produção de ouro, que é o segundo mineral mais

extraído no Brasil. A correlação que há desse *Cluster* com o *Cluster* 5 e 2 é pelo uso de palavras que se repetem nesses *Clusters*, como vale, minério e ferro. Cerca de 51,55% das notícias são dos anos de 2019, 2020 e 2021, indicando ser influenciado por notícias mais recentes e consequentemente pelos anos com maior quantidade de notícias. Foi denominado de *Anúncio da produção*.

O *Cluster* 5 é caracterizado pela quantidade de notícias, cerca de 1.441 ou 62,25% do total, sendo o maior *Cluster* em termo de quantidade de notícias. Ao analisar as palavras mais representativas desse *Cluster*, as notícias tratam do setor e mercado de mineração no Brasil, e da empresa mineradora Vale. Também aborda a questão de projeto, pois normalmente empresas de grande porte anunciam investimentos em projetos. Basicamente a maioria das notícias trazem essas palavras, o que justifica esse ser o maior *Cluster*. Em termos de quantidade de notícias por ano, cerca de 52,48% das notícias estão concentradas nos anos de 2019, 2020 e 2021, mostrando a influência das notícias dessas datas no *Cluster*, como também, esses serem os anos com maior quantidade de notícias. Foi denominado de *Júpiter*.

De forma geral, o contexto Sem Similaridade possui 52,48% das notícias concentradas nos anos de 2019, 2020 e 2021, mostrando que as notícias mais atuais possuem uma influência maior ao compor os *Clusters*. Pelo fato dos *Clusters* 2 e 5 terem a maior quantidade de notícias e tratarem de assuntos semelhantes possuem forte sobreposição, influenciando o resultado da pontuação *Silhouette*. Os *Clusters* 1, 3 e 4 por estarem próximos dos *Clusters* 2 e 5 guardam semelhança com eles, porém pouca entre si, indicando tratarem de assuntos diferentes.

Os *Clusters* tratam das variações da bolsa de valores de São Paulo e de Tóquio, das barragens de Minas Gerais, da empresa Vale, da produção de minério de ferro e de ouro e de projetos ligados a mineração, que possuem relevância para serem noticiados.

É possível captar a influência do tempo nos contextos, especialmente na análise dos *Clusters*, quando alguns são mais influenciados por notícias antigas, enquanto outros por notícias mais recentes. Outro fator importante a ser observado é a redução da quantidade de notícias, pois esse fato permite ao

algoritmo classificar de forma mais assertiva as notícias, porém, a sobreposição tende a manter. Esse fator pode ser justificado pelo tipo de notícias captadas, que apontam uma temática de bolsa de valores, produção de minério de ferro e novos projetos.

5 Conclusão

O objetivo dessa pesquisa foi desenvolver uma metodologia de recuperação da informação a partir da aplicação de sumarização automática de texto em notícias do setor de mineração no Brasil. A sumarização de texto é uma etapa importante na análise de inúmeros blocos de textos e assim obter *insights*. A aplicação da sumarização de texto nessa pesquisa permitiu a redução da quantidade de caracteres, cerca de 75%, possibilitando a inserção de outras técnicas de NLP. O uso de ATS permitiu aplicar o algoritmo de *Machine Learning* de *Clustering* para agrupar as notícias por padrão de palavras, e assim entender qual o discurso presente em cada agrupamento.

A aplicação desse ferramental permitiu identificar que há um efeito longitudinal, ou seja, a data da notícia influência na análise e, portanto, deve ser considerada. Quanto mais agregados os dados estiverem na análise, maior será esse efeito, pois um *cluster* pode ser composto na sua maioria por apenas um ou dois anos, aumentando a possibilidade da repetição de certos termos, uma vez que, a importância do discurso é dada pela repetição de termos.

A repetição de notícias com alto teor semântico tende a influenciar na construção dos *clusters*, mascarando informações relevantes e que devem ser mapeadas. As análises mostraram que ao retirar notícias com o mesmo teor semântico novas palavras surgem, trazendo a luz um assunto até então não abordado.

Dada as características das notícias, há uma tendência de alta sobreposição dos *clusters*, pois o valor da pontuação *silhoutte* tende a 0. Esse resultado mostra que muitas notícias estavam na borda de decisão, podendo ser classificada em um *cluster* ou em outro. O compartilhamento de termos pode ser uma das justificativas para ocorrer essa sobreposição. As notícias que se afastam

do núcleo do *Cluster*, fato esse mais perceptível nos *Clusters* 2 e 5, tendem a apresentar uma variação no discurso central do *Cluster*.

Outro fato observado na análise foram notícias relacionadas a projeto de ouro e o problema das barragens da empresa Vale em Minas Gerais. O método empregado também se mostrou eficaz em separar notícias que não estavam relacionadas a mineração de minerais. O *transformer* BERT também se mostrou eficaz em processar notícias no idioma inglês, e o método de *Clustering* foi eficiente por não misturar notícias do idioma inglês com o português.

A metodologia desenvolvida mostrou ser capaz de ser aplicada juntamente com ferramentas de análise, como 5 Forças de Porter, PEST, Matriz de Ansoff e outras, pois essas ferramentas demandam um grande volume textual para geração de sentido, principalmente em cada parte dos *frames* de análise. Essas ferramentas ajudam no ordenamento da informação possibilitando a análise e consequentemente a derivação de conclusões.

Por fim a metodologia e o processo desenvolvido mostrou ser capaz de processar multi-documentos, em um domínio específico para gerar uma saída sumarizada, aplicada ao setor de mineração. Nos anos com maior quantidade de notícias os dados mostram uma maior repetição semântica, com um discurso relacionado a variação da bolsa de São Paulo e produção de minério de ferro e ouro ao longo dos anos.

Referências

ABDEL-SALAM, Shehab; RAFEA, Ahmed. Performance study on extractive text summarization using BERT models. **Information**, Basel, v. 13, n. 2, p. 67, 2022. Disponível em: https://doi.org/10.3390/info13020067. Acesso em: 27 set. 2022.

ALAMI, Nabil; MEKNASSI, Mohammed; EN-NAHNAHI, Noureddine. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. **Expert Systems with Applications**, United Kingdom, v. 123, p. 195-211, 2019. Disponível em: https://doi.org/10.1016/j.eswa.2019.01.037. Acesso em: 27 jul. 2021.

AVERSA, Joseph; HERNANDEZ, Tony; DOHERTY, Sean. Incorporating big data within retail organizations: a case study approach. **Journal of Retailing and Consumer Services**, Amsterdan, v. 60, p. 1-9, 2021. Disponível em: https://doi.org/10.1016/j.jretconser.2021.102447. Acesso em: 6 abr. 2021.

BALDUINI, Marco *et al*. Models and practices in urban data science at scale. **Big Data Research**, Amsterdan, v. 17, p. 66-84, 2019. Disponível em: https://doi.org/10.1016/j.bdr.2018.04.003. Acesso em: 6 abr. 2021.

BONDIELLI, Alessandro; MARCELLONI, Francesco. On the use of summarization and transformer architectures for profiling résumés. **Expert Systems with Applications**, United Kingdom, v. 184, p. 1-10, 2021. Disponível em: https://doi.org/10.1016/j.eswa.2021.115521. Acesso em: 24 nov. 2021.

BRANDS, Kritine. Big data and business intelligence for management accountants. **Strategic Finance**, New Jersey, v. 95, p. 64-66, 2014.

CHOO, Chun Wei. **A organização do conhecimento:** como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo: SENAC, 2003.

CHOWDHURY, Gobinda G. Natural language processing. **Annual Review of Information Science and Technology**, New Jersey, v. 37, n. 1, p. 51-89, 2003. Disponível em: https://doi.org/10.1002/aris.1440370103. Acesso em: 21 mai. 2021.

CHRISTIAN, Hans; AGUS, Mikhael Pramodana; SUHARTONO, Derwin. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). **ComTech: Computer, Mathematics and Engineering Applications**, Jakarta, v. 7, n. 4, p. 285, 2016. Disponível em: https://doi.org/10.21512/comtech.v7i4.3746. Acesso em: 27 jul. 2021.

COPELAND, Michael. What's the difference between artificial intelligence, machine learning and deep learning? **NVIDIA**, [*s.l.*], 19 July 2016.

CRESWELL, John W; CRESWELL, J. David. **Research design: qualitative, quantitative, and mixed methods approaches**. 4. ed. Thousand Oaks,: Sage, 2014.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. **ArXiv**, Ithaca, v. 1, 2018. Disponível em: https://doi.org/10.48550/arXiv.1810.04805. Acesso em: 11 maio 2021.

DHAR, Vasant. Data science and prediction. **Communications of the ACM**, New York, v. 56, n. 12, p. 64-73, 2013. Disponível em: https://doi.org/10.1145/2500499. Acesso em: 27 set. 2022.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. Cambridge: MIT Press, 2016.

GOULARTE, Fábio Bif; NASSAR, Silvia Modesto; FILETO, Renato;

SAGGION, Horacio. A text summarization method based on fuzzy rules and applicable to automated assessment. **Expert Systems with Applications**, United Kingdom, v. 115, p. 264-275, 2019. Disponível em: https://doi.org/10.1016/j.eswa.2018.07.047. Acesso em: 16 abr. 2021.

GOULARTE, Fábio Bif; WILGES, Beatriz; NASSAR, Silvia Modesto; CISLAGHI, Renato. Métricas de sumarização automática de texto em tarefas de um ambiente virtual de aprendizagem. **Brazilian Symposium on Computers in Education**, Porto Alegre, p. 752, 2014. Disponível em: http:doi.org/10.5753/cbie.sbie.2014.752. Acesso em: 16 abr. 2021.
HAMET, Pavel; TREMBLAY, Johanne. Artificial intelligence in medicine. **Metabolism**, New York, v. 69, p. s36-s40, 2017. Disponível em: https://doi.org/10.1016/j.metabol.2017.01.011. Acesso em: 17 set. 2021.

HARK, Cengiz; KARCI, Ali. Karcı summarization: a simple and effective approach for automatic text summarization using Karcı entropy. **Information Processing & Management**, United Kingdom, v. 57, n. 3, p. 1-16, 2020. Disponível em: https://doi.org/10.1016/j.ipm.2019.102187. Acesso em: 27 jul. 2021.

JAIN, Priyank; GYANCHANDANI, Manasi; KHARE, Nilay. Big data privacy: a technological perspective and review. **Journal of Big Data**, Berlin, v. 3, n. 1, p. 25, 2016. Disponível em: http://doi.org/10.1186/s40537-016-0059-y. Acesso em: 6 abr. 2021.

JOHN, Ansamma; PREMJITH, P. S.; WILSCY, M. Extractive multi-document summarization using population-based multicriteria optimization. **Expert Systems with Applications**, United Kingdom, v. 86, p. 385-397, 2017. DOI: Disponível em: https://doi.org/10.1016/j.eswa.2017.05.075. Acesso em: 27 jul. 2021.

JOSHI, Akanksha; FIDALGO, E.; ALEGRE, E.; FERNÁNDEZ-ROBLES, Laura. SummCoder: an unsupervised framework for extractive text summarization based on deep auto-encoders. **Expert Systems with Applications**, United Kingdom, v. 129, p. 200-215, 2019. Disponível em: https://doi.org/10.1016/j.eswa.2019.03.045. Acesso em: 27 jul. 2021.

JOSHI, Aravind K. Natural language processing. **Science**, New York, v. 253, n. 5025, p. 1242-1249, 1991. Disponível em: https://doi.org/10.1126/science.253.5025.1242. Acesso em: 21 mai. 2021.

KHAMPARIA, Aditya; SINGH, Karan Mehtab. A systematic review on deep learning architectures and applications. **Expert Systems**, New Jersey, v. 36, n. 3, p. 1-22, 2019. Disponível em: https://doi.org/10.1111/exsy.12400. Acesso em: 11 nov. 2021.

LAMSIYAH, Salima; EL MAHDAOUY, Abdelkader; ESPINASSE, Bernard; EL ALAOUI OUATIK, Saïd. An unsupervised method for extractive multi-

document summarization based on centroid approach and sentence embeddings. **Expert Systems with Applications**, United Kingdom, v. 167, p. 114152, 2021a. Disponível em: https://doi.org/10.1016/j.eswa.2020.114152. Acesso em: 16 abr. 2021.

LAMSIYAH, Salima; MAHDAOUY, Abdelkader El; OUATIK, Saïd El Alaoui; ESPINASSE, Bernard. Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning. **Journal of Information Science**, United Kingdom, v. 49, n. 1, p. 164-182, 2021b. Disponível em: http://doi.org/10.1177/0165551521990616. Acesso em: 24 nov. 2021.

LEIJNEN, Stefan; VEEN, Fjodor Van. The neural network zoo. **Proceedings**, Basel, v. 47, n. 1, p. 9, 2020. Disponível em: https://doi.org/10.3390/proceedings2020047009**.** Acesso em: 05 nov. 2021.

LI, Ping; YU, Jiong. Extractive summarization based on dynamic memory network. **Symmetry**, Basel, v. 13, n. 4, p. 600, 2021. Disponível em: https://doi.org/10.3390/sym13040600. Acesso em: 27 set. 2022.

LIU, Yang; LAPATA, Mirella. Text summarization with pretrained encoders. **ArXiv**, Ithaca, v. 1, 2019. Disponível em: https://doi.org/10.48550/arXiv.1908.08345. Acesso em: 18 jul. 2022.

MCGEE, James; PRUSAK, Laurence. **Gerenciamento estratégico da informação**. Rio de Janeiro: Campus, 1994.

MILLER, Derek. Leveraging BERT for extractive text summarization on lectures. **ArXiv**, Ithaca, v. 1, 2019. Disponível em: http://doi.org/arxiv.org/abs/1906.04165. Acesso em: 27 ago. 2021.

MILLER, Jerry P. **O milênio da inteligência competitiva**. Porto Alegre: Bookman, 2002.

MUTLU, Begum; SEZER, Ebru A.; AKCAYOL, M. Ali. Candidate sentence selection for extractive text summarization. **Information Processing & Management**, United Kingdom, v. 57, n. 6, p. 1-18, 2020. Disponível em: https://doi.org/10.1016/j.ipm.2020.102359. Acesso em: 16 abr. 2021.

NESI, Paolo; PANTALEO, Gianni; SANESI, Gianmarco. A hadoop based platform for natural language processing of web pages and documents. **Journal of Visual Languages and Computing**, Amsterdan, v. 31, n. 2015, p. 130-138, 2015. Disponível em: http://doi.org/10.1016/j.jvlc.2015.10.017. Acesso em: 20 jun. 2020.

PADMAKUMAR, Aishwarya; SARAN, Akanksha. Unsupervised text summarization using sentence embeddings. **Tecnical Report**, University of Texas,Austin, p. 1-9, 2016.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

PROTIM GHOSH, Partha; SHAHARIAR, Rezvi; HOSSAIN KHAN, Muhammad Asif. A rule based extractive text summarization technique for bangla news documents. **International Journal of Modern Education and Computer Science**, Hong Kong, v. 10, n. 12, p. 44-53, 2018. Disponível em: http://doi.org/10.5815/ijmecs.2018.12.06. Acesso em: 27 jul. 2021.

RAMOS, Hélia de Sousa Chaves; BRÄSCHER, Marisa. Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T. **Ciência da Informação**, Brasília, v. 38, n. 2, p. 56-68, 2009. Disponível em: http://doi.org/10.1590/s0100-19652009000200005. Acesso em: 22 abr. 2021.

RICHARDSON, Roberto Jarry. **Pesquisa Social**: métodos e técnicas. 3. ed. São Paulo: Atlas, 2012.

RINALDI, Antonio M.; RUSSO, Cristiano; TOMMASINO, Cristian. A semantic approach for document classification using deep neural networks and multimedia knowledge graph. **Expert Systems with Applications**, United Kingdom, v. 169, p. 1-13, 2021. Disponível em: http://doi.org/10.1016/j.eswa.2020.114320. Acesso em: 14 abr. 2021.

SALEHI, Hadi; BURGUEÑO, Rigoberto. Emerging artificial intelligence methods in structural engineering. **Engineering Structures**, United Kingdom, v. 171, p. 170-189, 2018. Disponível em: https://doi.org/10.1016/j.engstruct.2018.05.084. Acesso em: 17 set. 2021.

SEARLE, Thomas; IBRAHIM, Zina; TEO, James; DOBSON, Richard JB. Estimating redundancy in clinical text. **ArXiv**, Ithaca, v. 1, 2021. Disponível em: https://doi.org/10.48550/arXiv.2105.11832. Acesso em: 24 nov. 2021.

SHRESTHA, Ajay; MAHMOOD, Ausif. Review of deep learning algorithms and architectures. **IEEE Access**, New York, v. 7, p. 53040-53065, 2019. Disponível em: https://di.org/10.1109/ACCESS.2019.2912200. Acesso em: 11 nov. 2021.

SINHA, Aakash; YADAV, Abhishek; GAHLOT, Akshay. Extractive text summarization using neural networks. **ArXiv**, Ithaca, v. 1, 2018. Disponível em: https://doi.org/10.48550/arXiv.1802.10137. Acesso em: 27 ago. 2021.

STANTON, Jeffrey M. Data science: what's in it for the new librarian? **Syracuse University**, New York, 26 July 2012.

SYED, Ayesha Ayub; GAOL, Ford Lumban; MATSUO, Tokuro. A survey of

the state-of-the-art models in neural abstractive text summarization. **IEEE Access**, New York, v. 9, p. 13248-13265, 2021. Disponível em:https://doi.org/10.1109/ACCESS.2021.3052783. Acesso em: 16 abr. 2021.

TAN, Bowen; KIEUVONGNGAM, Virapat; NIU, Yiming. Automatic text summarization of covid-19 medical research articles using BERT and GPT-2. **ArXiv**, Ithaca, v. 1, 2020. Disponível em: https://doi.org/10.48550/arXiv.2006.01997. Acesso em: 17 abr. 2021.

VASCONCELLOS, Vera M. Ramos; SILVA, Anne P. P. Nascimento; SOUZA, Roberta Teixeira. O estado da arte ou o estado do conhecimento. **Educação**, Porto Alegre, v. 43, n. 3, p. 1-12, 2020. Disponível em: htttps://doi.org/10.15448/1981-2582.2020.3.37452. Acesso em: 26 nov. 2021.

WANG, Lin. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, United Kingdom, v. 74, n. 6, p. 1243-1257, 2018. Disponível em: https://doi.org/10.1108/JD-02-2018-0036. Acesso em: 27 set. 2022.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. 2. ed. Rio de Janeiro: LTC, 2014.

WEAVER, Adam. Tourism, big data, and a crisis of analysis. **Annals of Tourism Research**, United Kingdom, v. 88, p. 1-11, 2021. Disponível em: https://doi.org/10.1016/j.annals.2021.103158. Acesso em: 16 abr. 2021.

WOLF, Thomas *et al*. HuggingFace's transformers: state-of-the-art natural language processing. **ArXiv**, Ithaca, v. 1, 2019. Disponível em: https://doi.org/10.48550/arXiv.1910.03771. Acesso em: 11 nov. 2021.

YANG, Guangbing *et al*. The effectiveness of automatic text summarization in mobile learning contexts. **Computers & Education**, Amsterdan, v. 68, p. 233-243, 2013. Disponível em: https://doi.org/10.1016/j.compedu.2013.05.012. Acesso em: 19 abr. 2021.

ZHANG, Aston; LIPTON, Zachary C.; LI, Mu; SMOLA, Alexander J. **Dive into deep learning**. Cambridge: Cambridge University Press; 2020.

Use of deep learning to build an information retrieval model applied to the mining sector in Brazil

Abstract: Faced with the exponential growth of data and information, provided by sensors and social media, an ecosystem composed of new storage and processing infrastructures, called Big Data, was developed. All this development resulted in a new area of knowledge, called Data Science. Despite there being an

ecosystem and an area of knowledge to deal with this massive block of data and information, the discomfort of an overabundance of data still remains and becomes more significant when companies become aware that they can use zettabytes of data and information to direct their strategy and operations. Based on this, this research sought to develop a method to summarize news from the mining sector in Brazil, identifying the effect of semantic similarity in the analysis, enabling information retrieval and use in processes of understanding the sector. In this method, the BERTSUM transformer was applied to summarize the news, and after summarizing, the BERT transformer was applied to measure the similarity between the news. The method made it possible to reduce the entire block of text by 75%, remove news with the same semantic content, and deduce that there is a pattern in the discourse of news related to the mining sector.

Keywords: natural language processing ; deep learning; BERT; ATS; mining

Recebido: 14/09/2023
Aceito: 22/11/2023

Declaração de autoria:

Concepção e elaboração do estudo: Luander Falcão, Brenner Lopes, Renato Souza e Ricardo Barbosa.
Coleta de dados: Luander Falcão, Brenner Lopes, Renato Souza e Ricardo Barbosa.
Análise e interpretação dos dados: Luander Falcão, Brenner Lopes, Renato Souza e Ricardo Barbosa.
Redação: Luander Falcão, Brenner Lopes, Renato Souza e Ricardo Barbosa.
Revisão crítica do manuscrito: Luander Falcão, Brenner Lopes, Renato Souza e Ricardo Barbosa.

Como citar

FALCÃO, Luander Cipriano de Jesus; LOPES, Brenner; SOUZA, Renato Rocha; BARBOSA, Ricardo Rodrigues. Uso de deep learning para a construção de um modelo de recuperação da informação aplicado para o setor de mineração no Brasil. **Em Questão**, Porto Alegre, v. 30, e-135550, 2024. DOI: https://doi.org/10.1590/1808-5245.30.135550