UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO

LUANDER CIPRIANO DE JESUS FALCÃO

SUMARIZAÇÃO DE TEXTO EM DEEP LEARNING COMO ETAPA INICIAL PARA A CONSTRUÇÃO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO: ANÁLISE DO SETOR DE MINERAÇÃO NO BRASIL

Belo Horizonte
2024

LUANDER CIPRIANO DE JESUS FALCÃO

SUMARIZAÇÃO DE TEXTO EM DEEP LEARNING COMO ETAPA INICIAL PARA A CONSTRUÇÃO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO: ANÁLISE DO SETOR DE MINERAÇÃO NO BRASIL

Tese apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação da Universidade Federal de Minas Gerais para obtenção do grau de Doutor, área de concentração Ciência da Informação.

Linha de Pesquisa: Gestão e Tecnologia da Informação e Comunicação

Orientador: Renato Rocha Souza

BELO HORIZONTE

2024

F178s Falcão, Luander Cipriano de Jesus.
Sumarização de texto em Deep Learning como etapa inicial para a construção de um modelo de recuperação da informação [recurso eletrônico] : análise do setor de mineração no Brasil / Luander Cipriano de Jesus Falcão. - 2024.
1 recurso online (104f. : il., color.) : pdf.

Orientador: Renato Rocha Souza.
Tese (doutorado)ï Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 95-104.
Exigência do sistema: Adobe Acrobat Reader.

1. Ciência da informação ï Teses. 2. Recuperação da informação - Teses. 3. Resumos - redação - Teses. 4. Inteligência artificial ï processamento de dados ï Teses. 5. Minas e recursos minerais - Teses. I. Souza, Renato Rocha. II. Universidade Federal de Minas Gerais. Escola de Ciência da Informação. III. Título.

CDU: 025.4.03

Ficha catalográfica: Maianna Giselle de Paula ï CRB6: 2642

Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

FOLHA DE APROVAÇÃO

SUMARIZAÇÃO DE TEXTO EM DEEP LEARNING COMO ETAPA INICIAL PARA A CONSTRUÇÃO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO: ANÁLISE DO SETOR DE MINERAÇÃO NO BRASIL

LUANDER CIPRIANO DE JESUS FALCÃO

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia da Informação e Comunicação.

Aprovada em 13 de maio de 2024, por videoconferência, pela banca constituída pelos membros:

Prof(a). Renato Rocha Souza (Orientador)
FGV/RJ

Prof(a). Carlos Henrique Marcondes de Almeida
UFF

Prof(a). Frederico Cesar Mafra Pereira
ECI/UFMG

Prof(a). George Leal Jamil
Fundação Dom Cabral

Prof(a). Gustavo Quiroga Souki
Instituto Superior Manuel Teixeira Gomes, Portugal

Prof(a). Ricardo Rodrigues Barbosa
Aposentado/UFMG

Belo Horizonte, 13 de maio de 2024.

Documento assinado eletronicamente por **Renato Rocha Souza**, **Usuário Externo**, em 15/05/2024, às 03:51, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **George Leal Jamil**, **Usuário Externo**, em 15/05/2024, às 19:22, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Henrique Marcondes de Almeida**, **Professor do Magistério Superior - Visitante**, em 19/06/2024, às 17:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Frederico Cesar Mafra Pereira**, **Professor do Magistério Superior**, em 19/06/2024, às 18:23, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gustavo Quiroga Souki**, **Usuário Externo**, em 26/06/2024, às 05:21, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Rodrigues Barbosa**, **Membro de comissão**, em 03/07/2024, às 20:19, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3241520** e o código CRC **F9F47E72**.

Referência: Processo nº 23072.225753/2024-63 SEI nº 3241520

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

ATA DA DEFESA DE TESE DO ALUNO

LUANDER CIPRIANO DE JESUS FALCÃO

Realizou-se, no dia 13 de maio de 2024, às 14:00 horas, por videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *SUMARIZAÇÃO DE TEXTO EM DEEP LEARNING COMO ETAPA INICIAL PARA A CONSTRUÇÃO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO: ANÁLISE DO SETOR DE MINERAÇÃO NO BRASIL*, apresentada por LUANDER CIPRIANO DE JESUS FALCÃO, número de registro 2020660673, graduado no curso de CIÊNCIAS ECONÔMICAS, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Renato Rocha Souza - FGV/RJ (Orientador), Prof(a). Carlos Henrique Marcondes de Almeida - UFF, Prof(a). Frederico Cesar Mafra Pereira - ECI/UFMG, Prof(a). George Leal Jamil - Fundação Dom Cabral, Prof(a). Gustavo Quiroga Souki - Instituto Superior Manuel Teixeira Gomes, Portugal, Prof(a). Ricardo Rodrigues Barbosa - Aposentado/UFMG.

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 13 de maio de 2024.

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Renato Rocha Souza**, **Usuário Externo**, em 15/05/2024, às 03:51, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **George Leal Jamil**, **Usuário Externo**, em 15/05/2024, às 19:22, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Henrique Marcondes de Almeida**, **Professor do Magistério Superior - Visitante**, em 19/06/2024, às 17:38, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Frederico Cesar Mafra Pereira**, **Professor do Magistério Superior**, em 19/06/2024, às 18:23, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gustavo Quiroga Souki**, **Usuário Externo**, em 26/06/2024, às 05:21, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Rodrigues Barbosa**, **Membro de comissão**, em 03/07/2024, às 20:19, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3241518** e o código CRC **E9F297F3**.

Referência: Processo nº 23072.225753/2024-63 SEI nº 3241518

DEDICATÓRIA

Ao Deus Todo-Poderoso, por sua eterna graça e misericórdia sobre a minha vida.

À minha amada esposa, Quenia, e aos meus filhos, Gabriel e Beatriz.

AGRADECIMENTOS

À minha amada esposa Quenia, pela paciência e pelo companheirismo durante todos esses anos, sempre me encorajando e confiando em mim. Aos meus filhos, que, mesmo sem entender, sempre acreditaram. Vocês resumem tudo de mais precioso que tenho na vida.

Aos meus pais e às minhas irmãs, Luana e Taiana.

Ao amigo Brenner Lopes, que também esteve comigo nessa jornada.

Ao Prof. Renato Souza, pela confiança e orientação ao longo dessa caminhada.

A todos os familiares, amigos e colegas que me ajudaram em alguma etapa do doutorado: o meu muito obrigado.

“Não havendo sábia direção, cai o povo,
mas na multidão de conselheiros há
segurança.” (Provérbios 11.14)

RESUMO

Na virada do século XX para o XXI, o mundo experimentou uma mudança de paradigma, saindo de um estado de escassez de dados para um estado de superabundância de dados. Esse novo cenário gerou o *Big Data*, uma série de ferramentas e de tecnologias próprias para o tratamento e o armazenamento de dados. Atrelada a ele está a Inteligência Artificial, que ganha mais relevância por proporcionar novos métodos para tratar enormes silos de dados, inclusive textuais, por meio do uso de *Natural Language Processing*. Apesar de o *Natural Language Processing* possuir várias tarefas, o Resumo Automático de Texto se destaca por reduzir a quantidade de escrita sem perder o sentido, proporcionando a aplicação de outras tarefas de *Natural Language Processing* e de algoritmos de *Machine Learning*. Diante desse cenário, surge a necessidade de recuperar e de analisar informações que mostrem mudanças estruturais no macroambiente do setor de mineração no Brasil. Para isso, foram coletadas, pré-processadas e sumarizadas cerca de 3.224 notícias de *sites* e de jornais sobre esse setor. Após sumarizadas, as notícias passaram por etapas de medição de similaridade, de mensuração do grau de sentimento e de *clustering* dos resumos. Os resumos, por sua vez, foram agrupados em contextos Geral e Sem Similaridade Semântica, e para cada ano do contexto Sem Similaridade Semântica também foram gerados *Clusters* para análise. Na sequência, foram construídos um *dataframe,* com o *out-put* final, e um painel de dados. Em seguida, o método construído foi avaliado por 15 especialistas, sendo 4 de Mineração e 11 de Dados e Informação, dos quais 2 possuíam especialização em ambas as áreas. Em termos de resultados, a aplicação de todo esse ferramental permitiu identificar que há um efeito longitudinal nos dados. 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 à luz um assunto até então não abordado. A maioria das notícias utiliza o mesmo agrupamento de palavras. Esse agrupamento ocorre devido à repetição das palavras e ao fato de essas palavras estarem em dois ou três *Clusters*. A metodologia desenvolvida evidencia a capacidade de aplicação juntamente com as técnicas envolvidas na análise de negócios, nos dados competitivos e nas informações, tanto as clássicas quanto as contemporâneas mais populares, da inteligência e da estratégia competitiva.

Palavras-chave: Resumo automático de texto (ATS); BERT; *Deep Learning;* Macroambiente; Mineração.

ABSTRACT

At the turn of the 20th century to the 21st, the world experienced a paradigm shift, moving from a state of data scarcity to a state of data overabundance. This new scenario generated Big Data, a series of tools and technologies for processing and storing data. Linked to it is Artificial Intelligence, which gains more relevance by providing new methods for dealing with huge silos of data, including textual ones, through the use of Natural Language Processing. Although Natural Language Processing has several tasks, Automatic Text Summary stands out for reducing the amount of writing without losing meaning, providing the application of other Natural Language Processing tasks and Machine Learning algorithms. Given this scenario, there is a need to retrieve and analyze information that shows structural changes in the macroenvironment of the mining sector in Brazil. To this end, around 3,224 news items from websites and newspapers about this sector were collected, pre-processed and summarized. After being summarized, the news went through steps of measuring similarity, measuring the degree of sentiment and clustering the summaries. The summaries, in turn, were grouped into General and No Semantic Similarity contexts, and for each year of the No Semantic Similarity context, Clusters were also generated for analysis. Next, a dataframe was built, with the final output, and a data panel. Then, the constructed method was evaluated by 15 experts, 4 from Mining and 11 from Data and Information, of which 2 had specialization in both areas. In terms of results, the application of all this tooling allowed us to identify that there is a longitudinal effect in the data. The repetition of news with a high semantic content tends to influence the construction of Clusters, masking relevant information that must be mapped. The analyzes showed that when removing news with the same semantic content, new words emerge, bringing to light a subject that had not been covered until then. Most news stories use the same grouping of words. This grouping occurs due to the repetition of words and the fact that these words are in two or three Clusters. The methodology developed highlights the ability to apply together with the techniques involved in business analysis, competitive data and information, both classic and the most popular contemporary ones, intelligence and competitive strategy.

Keywords: Automated text summarization (ATS); BERT; Deep Learning; Macroenvironment; Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

ABIMAQ – ASSOCIAÇÃO BRASILEIRA DE MÁQUINAS

AI – ARTIFICIAL INTELLIGENCE

ANM – AGÊNCIA NACIONAL DE MINERAÇÃO

ATS – AUTOMATED TEXT SUMMARIZATION

BERT – BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

BTS – BIOMEDICAL TEXT SUMMARY

CBOW – THE CONTINUOUS BAG OF WORDS MODEL

CFEM – CONTRIBUIÇÃO FINANCEIRA PELA EXPLORAÇÃO MINERAL

CNN – CONVOLUTIONAL NEURAL NETWORKS

COVE – CONTEXT VECTORS

CPRM – SERVIÇO GEOLÓGICO DO BRASIL

DL – *DEEP LEARNING*

DNN – DEEP NEURAL NETWORKS

EHR – ELECTRONIC HEALTH RECORDS

ELMO – EMBEDDINGS FROM LANGUAGE MODELS

GAN – GENERATIVE ADVERSARIAL NETWORKS

GLOVE – WORD EMBEDDING WITH GLOBAL VECTORS

GPT – GENERATIVE PRE-TRAINED TRANSFORMERS

GPU – GRAPHICS PROCESSING UNIT

GRU – GATED RECURRENT UNITS

IBOVESPA – BOLSA DE VALORES DE SÃO PAULO

IOT – INTERNET DAS COISAS

LLM – LARGE LANGUAGE MODELS

LSTM – LONG SHORT-TERM MEMORY

ML – MACHINE LEARNING

MME – MINISTÉRIO DE MINAS E ENERGIA

NLM – NEURAL LANGUAGE MODELS

NLP – NATURAL LANGUAGE PROCESSING

NTM – NEURAL TURING MACHINE

PCA – PRINCIPAL COMPONENT ANALYSIS

PLN – PROCESSAMENTO DE LINGUAGEM NATURAL

PPI – PROGRAMA DE PARCERIAS DE INVESTIMENTOS

RNN – RECURRENT NEURAL NETWORKS

TF-IDF – Term Frequency – Inverse Data Frequency

WMC – WORLD, MINING, CONGRESS

WORD2VEC – WORD EMBEDDING

SUMÁRIO

1. INTRODUÇÃO

Neste capítulo são apresentadas as motivações e as lacunas encontradas que justificam o desenvolvimento desse trabalho de pesquisa. Na primeira seção (1.1 – Problema de Pesquisa) é apresentado o problema, que está relacionado à alta disponibilidade de dados e informações e a como transformar essas massivas bases em conhecimento acionável. Na segunda seção (1.2 – Justificativa) são apresentados os motivos que tornam essa pesquisa relevante. Na sequência, na terceira e na quarta seções, são apresentados o Objetivo Geral e os Objetivos Específicos, respectivamente, os quais nortearão as etapas tanto de referencial teórico quanto de metodologia, que essa pesquisa percorre com o foco de responder à pergunta central desse projeto.

1.1 Problema de Pesquisa

Em um mundo de dados e de informações produzidos, armazenados, recuperados e compartilhados com maior velocidade a cada ano, a disponibilidade de acesso é eminente. Essa disponibilidade é gerada pelo crescimento exponencial da instalação de sensores, do poder computacional, das redes sociais, de IoT (Internet das Coisas) 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). A esse fenômeno da explosão de dados e de informações dá-se o nome de *Big Data*.

Apesar de ser algo novo e de haver várias definições, há uma convergência de entendimento no sentido de o *Big Data* classificar-se como “especificamente um conjunto de dados tão grandes ou complexos que os aplicativos tradicionais de processamento de dados não são suficientes” (Jain; Gyanchandani; Khare, 2016, p. 1, tradução nossa) para registrar e para 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 de recuperação da informação por meio das tecnologias atreladas ao *Big Data*, como a universalização da Inteligência Artificial (*Artificial Intelligence – AI*), e às aplicações das técnicas analíticas no campo do *Machine Learning* (ML) e *Deep Learning* (DL), originou-se um novo campo de estudo, denominado Ciência de Dados.

Diante dessa confluência de várias tendências importantes na ciência, incluindo a prevalência de *Big Data*, o desenvolvimento de abordagens computacionais para análise e a necessidade de reprodutibilidade na investigação se tornaram os principais componentes da ciência de dados (Goldsmith *et al*., 2021). O conceito de Ciência de Dados é dado como uma nova série de tecnologias, de processos e de sistemas para extrair valor e fazer descobertas (Wang, 2018), tendo como missão a transformação de dados brutos e confusos em um conhecimento que possa ser aplicável (Stanton, 2012).

Para Dhar (2013), Ciência de Dados é o estudo sistemático da organização, das propriedades e da análise de dados e de seu papel na inferência. Esse conceito traz uma visão mais pragmática para o *Big Data* por focar na coleta, na preparação, na análise, na visualização, no gerenciamento e na preservação de grandes coleções de dados, não se limitando apenas aos conhecimentos inerentes à ciência da computação, mas tendo interface com a matemática, a estatística e com outras áreas do conhecimento. Diante dessas características, a Ciência de Dados possui uma maior capacidade de descobrir percepções e tendências valiosas e desconhecidas em um conjunto de dados de uma área onde o conhecimento é limitado (Salehi; Burgueño, 2018), com foco na produção de *insights* analíticos e no estabelecimento de modelos de previsão (Wang, 2018).

Um dos pontos centrais de todo esse contexto envolve a recuperação da informação. Para a recuperação da informação é necessária a construção de arquiteturas de dados acessados via programas de computador, “que ajudam as pessoas a acessar grandes coleções e encontrar dentro dessas coleções itens de interesse” (Furner, 2015, p. 3, tradução nossa). No caso de empresas, independentemente do ramo de negócios, a recuperação da informação está ligada ao provimento de 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 grandes bancos de dados conectados e disponíveis para gerar valor, surge o incômodo problema da sobrecarga de informações. Esse incômodo é potencializado quando se entende que é possível obter vantagem competitiva a partir de documentos textuais ou de informação textual em formato digital, que representam 80% da informação que circula na *WEB* (Lamsiyah *et al*., 2021b). Porém, esse é um processo complexo, exaustivo e caro quando se utiliza recursos humanos, mesmo havendo um enorme silo de informação disponível para identificação e extração de ideias centrais e úteis dos textos (Mutlu; Sezer; Akcayol, 2020; Yang *et al*., 2013).

Mesmo com toda evolução e disponibilização tecnológica, a recuperação de dados pode ser afetada pela representação e pela organização das informações (Weissenberger, 2015), gerando blocos de informações sem utilidade. Por isso, foram

desenvolvidos métodos computacionais focados em trabalhar com textos que visam reduzir a dimensão dos dados brutos para economizar tempo, recursos e encontrar as informações mais adequadas às necessidades do usuário (Lamsiyah *et al*., 2021b; Mutlu; Sezer; Akcayol, 2020).

Sendo assim, todo esse fluxo de inovações tecnológicas tende a continuar, “pois mais e mais trabalhadores passarão da economia física para a economia da informação, e as pessoas passarão mais tempo de trabalho e lazer criando, manipulando e comunicando informações” (Snir, 2011, p. 38, tradução nossa). O resultado mais proeminente dessa previsão foi a migração de um estado de escassez de dados para um estado de superabundância. Empresas de todos os setores produtivos, tanto as tradicionais quanto as de tecnologia, passaram a utilizar essa superabundância para direcionar suas atividades, cada vez mais centradas em dados. Com a análise desses dados, as empresas buscam entender melhor seus mercados e tomar decisões aproveitando as oportunidades (Aversa; Hernandez; Doherty, 2021). A finalidade dessas ações é “criar *insights* acionáveis para a entrega sustentada de valor, medir o desempenho e estabelecer vantagens competitivas” (Fosso Wamba *et al*., 2015, p. 235, tradução nossa).

1.2 Justificativa

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 de negócio que orientam o uso das informações (Weaver, 2021). No entanto, a resolução desses desafios esbarra na carência de mão de obra qualificada para trabalhar no ecossistema do *Big Data*. Faltam cientistas especializados para aplicar os métodos relacionados à Ciência de Dados, como a coleta, a preparação, a análise, a visualização, o gerenciamento e a preservação de grandes coleções de dados. Há, também, a falta de tempo para interpretar os dados resultantes das tecnologias de *Big Data* e a incapacidade humana de assimilar grandes quantidades de informações, inclusive na forma de dados não estruturados e de cunho textual, tornando métodos de resumo de texto elementos eficientes e importantes (Padmakumar; Saran, 2016).

O método de sumarização de texto apresenta fatores superiores quando comparado a métodos como indexação de palavras-chaves e catalogação de artigos por captar a semântica do texto, logo, sendo capaz de recuperar uma informação mais pertinente à necessidade do usuário. Com a sumarização de texto, é possível resumir um

conteúdo longo, seja no formato de notícias, seja em outro formato, em uma versão mais curta (Protim Ghosh; Shahariar; Hossain Khan, 2018).

A sumarização de texto pode ser aplicada às notícias de qualquer setor empresarial, principalmente quando esse é muito 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 que quanto maior for a quantidade de notícias, maior será a intensidade informacional de um setor, e, consequentemente, mais dinâmico ele será.

Um setor no Brasil que possui essa característica de dinamicidade é o de Mineração, um setor de alta importância no país, pois corresponde a, aproximadamente, 4% do PIB do Brasil. A Mineração emprega mais de 170 mil empregados diretos e o Brasil é exportador global de Nióbio (1º lugar), Ferro (2º lugar), Vermiculita (3º lugar) e Grafita (3º lugar) (Ibram, 2021). No ano de 2020, o faturamento do setor foi de R$ 83 bi e foram gerados R$ 28,8 bi de arrecadação de impostos, incluindo CFEM (Contribuição Financeira pela Exploração Mineral). As principais substâncias produzidas no Brasil, em termos de participação no faturamento do setor, são Minério de Ferro, Minério de Ouro e Minério de Cobre, que correspondem a 73,51%, 8,74% e 6,31%, respectivamente, do total (Ibram, 2021).

Partindo da premissa de que em “todos os segmentos de negócio, uma empresa chegará à liderança através do uso da informação como uma vantagem competitiva, e no processo, mudando as regras da competição para todo mundo” (Mcgee; Prusak, 1994, p. 71), explica-se o uso intensivo de Ciência de Dados. Ao considerar a importância do setor de Mineração, com toda a sua ramificação no Brasil, e o fato de haver grandes repositórios textuais com potencial de serem transformados em informações preciosas para a organização (Ramos; Bräscher, 2009), justifica-se a aplicação das modernas ferramentas de *Big Data* e de Ciência de Dados para tratamento de texto.

Dada a importância do setor e, consequentemente, a vasta produção de material textual em torno dele, principalmente de notícias, faz-se com que ele seja dinâmico. As notícias de um setor contam o que está acontecendo. 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 macroambiente e seus respectivos aspectos.

O macroambiente é o nível do ambiente de uma organização e está além da influência direta ou do controle primário de qualquer organização (Fleisher; Bensoussan,

2003). Para Porter (2005, p. 22), o macroambiente compreende a dimensão externa às organizações e às empresas, sendo que “as forças externas, em geral, afetam todas as empresas na indústria, e o ponto básico encontra-se nas diferentes habilidades das empresas em lidar com elas”. De acordo com Fleisher e Bensoussan (2003), o macroambiente possui a capacidade de impactar a competitividade tanto do setor como da empresa, sendo necessário monitorá-lo, pois ele aponta as mudanças e as tendências tanto do setor quanto das empresas, as quais devem se moldar e adaptar suas estratégias. Um exemplo das possíveis mudanças seria a alteração da taxa de juros ou oscilações cambiais, ou a promulgação de novas leis. Em termos de tendência a mudança de comportamento da população, em relação a consumo de certos bens e serviços, ou o envelhecimento da população, gerando uma demanda por novos tipos de negócios.

Segundo Fleisher e Bensoussan (2003), o macroambiente pode ser monitorado e analisado por meio das dimensões de ordem social, tecnológica, econômica, ecológica/ambiental, cultural e político/legal/regulatória. Essa estrutura de composição é a mais comum na literatura para analisá-lo (Comai; Millán, 2006), e essas dimensões possibilitam a criação de sentido quanto ao atual contexto do setor e à qual a tendência será a adotada.

Diante dessa complexidade, surge a necessidade de construção de modelos capazes de recuperar informações mais aderentes à necessidade de quem demanda a informação e de apresentá-las de forma eficiente e eficaz.

Nesse contexto, o objetivo de pesquisa deste trabalho é tratar um grande volume de informação textual no formato de notícia. Essas notícias são artigos de jornais e de *press releases*, provenientes de *sites* especializados em mineração e de jornais de abrangência municipal, estadual e nacional. Todas as notícias coletadas passarão pela etapa inicial: a sumarização de texto, para gerar um novo patamar de compreensão do Setor de Mineração no Brasil.

Tendo exposto este contexto, a pergunta central que esse trabalho pretende responder é: “Como recuperar informações que mostrem mudanças estruturais no macroambiente do setor de mineração no Brasil com a aplicação de sumarização de texto implementada por *Deep Learning*?”

1.3 Objetivo geral

O objetivo deste estudo é propor uma metodologia de recuperação da informação a partir do uso de sumarização automática de texto para monitorar a evolução de mudanças estruturais no macroambiente do setor de mineração no Brasil.

1.4 Objetivos específicos

São os objetivos específicos que sustentarão o objetivo geral:

- Criar uma rotina de sumarização de texto, gerando resumos das notícias coletadas;
- Medir a similaridade textual dos resumos para quantificar quantos deles possuem o mesmo teor semântico para evitar duplicidades semânticas;
- Classificar os resumos quanto ao seu grau de sentimento, sendo eles: positivo, neutro ou negativo;
- Criar um modelo de visualização da informação para explorar e para suportar a obtenção de *insights* relevantes de entendimento e de criação de sentido;
- Mensurar a repetição de narrativas/discursos ao longo do tempo;
- Submeter o modelo de visualização e os *insights* encontrados a uma avaliação de especialistas do setor de mineração e de especialistas em Dados e Informação para entender o quanto essa metodologia é capaz de mudar a compreensão e de criar sentido para eles;

1.5 Estrutura da tese

A tese está estruturada em seis capítulos, conforme ilustrado na Figura 1:

FIGURA 1: Estrutura da Tese

Fonte: elaborado pelo autor.

Na presente introdução, foram apresentadas as motivações e as lacunas encontradas para justificar o desenvolvimento desse trabalho de pesquisa. A problemática é apresentada em detalhes e relacionada aos objetivos desta pesquisa.

No capítulo 2, são apresentados dois blocos, sendo um sobre os marcos teóricos e outro sobre o estado da arte. O bloco de marcos teóricos compreende os temas necessários para embasar o entendimento deste trabalho, como os conceitos de Resumo Automático de Texto e de *Deep Learning*. O bloco sobre o estado da arte compreende as pesquisas mais recentes que trabalharam Resumo Automático de Texto com *Deep Learning*.

No capítulo 3, são apresentados, em detalhes, a metodologia, tanto de construção quanto de validação do modelo, e os recursos computacionais utilizados em todos os testes empíricos. O capítulo 4 relata os resultados do experimento e o capítulo 5, a validação desse experimento. No capítulo 6, são desenvolvidas as conclusões sobre os resultados obtidos a partir da metodologia e as considerações finais sobre as futuras potenciais pesquisas oriundas dessa tese. Por fim, as referências bibliográficas relevantes são enunciadas.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

Neste capítulo, as seções foram agrupadas em dois blocos, sendo um a respeito de marcos teóricos e outro sobre o estado da arte. O bloco de marcos teóricos compreende todas as seções necessárias para embasar o entendimento deste trabalho, e a seção sobre o estado da arte compreende as pesquisas mais recentes que tratam do objetivo desse trabalho ou que trabalham com Resumo Automático de Texto com *Deep Learning*.

O bloco de marcos teóricos é composto por três seções. Na primeira, são apresentados os fundamentos do resumo automático de texto, com o objetivo de compreender como surge essa temática e como ela se desenvolveu, para enfatizar a diferença entre metodologias coirmãs.

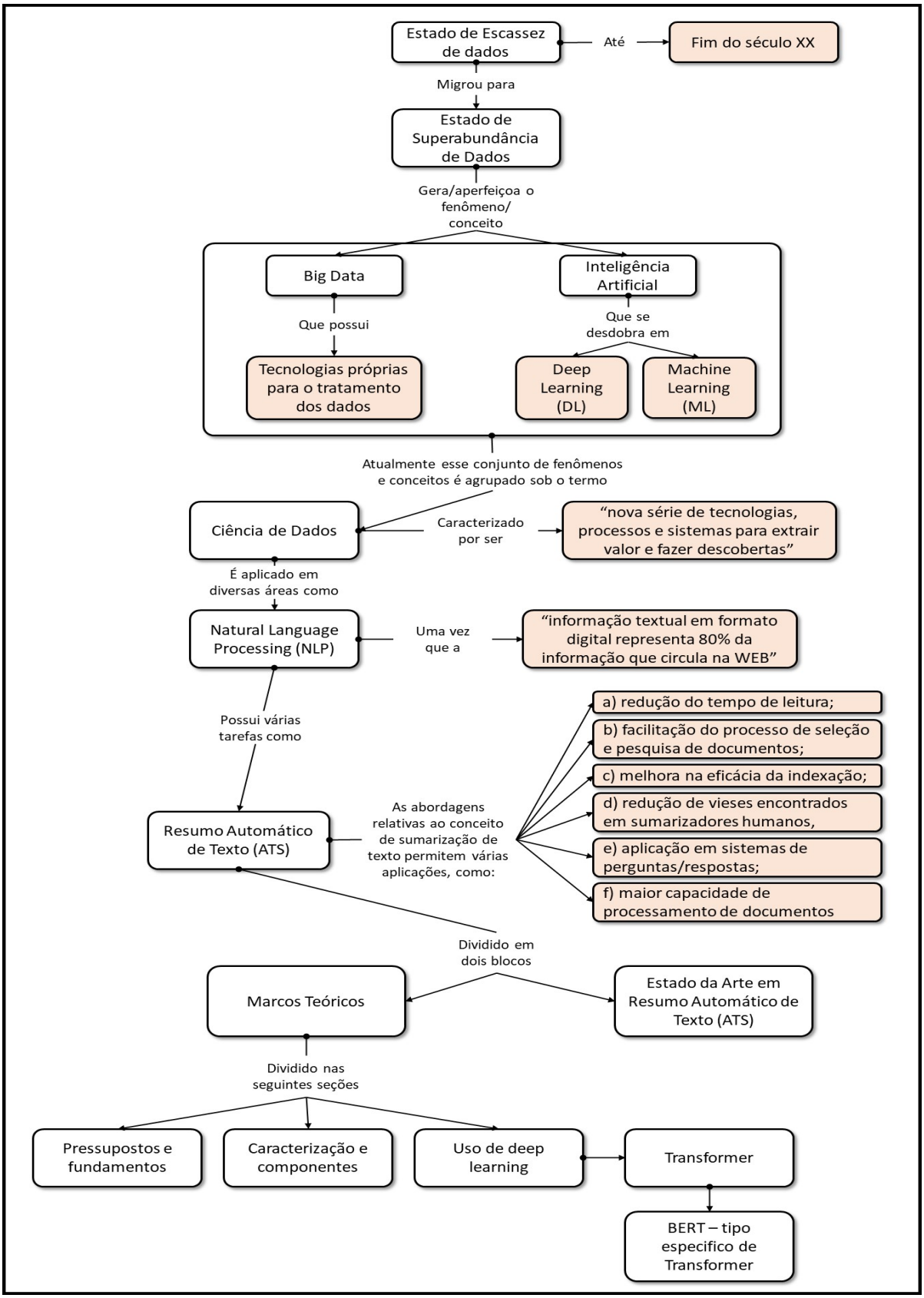
Na segunda seção, são apresentadas as principais características do Resumo Automático de Texto, contemplando os principais termos e definições utilizados nessa temática, e as principais características quando a base consiste no tipo de saída, no tipo de entrada e no propósito desse resumo.

Na terceira seção, é apresentado o uso de *Deep Learning* para resumo de texto. Entre os motivos dos algoritmos de *Deep Learning* serem melhores do que os algoritmos de *Machine Learning* está o fato de usarem Redes Neurais com várias camadas. O rápido aperfeiçoamento das redes neurais de múltiplas camadas em *Deep Learning* permitiu o surgimento de novas arquiteturas capazes de tratar, de forma mais eficiente, sons, imagens, vídeos e, inclusive, textos. Isso deu origem a uma série de algoritmos, no caso de *Deep Learning*, também chamados de Arquiteturas, voltadas para executarem tarefas específicas de *Natural Language Processing* (NLP). As arquiteturas próprias de *Deep Learning* direcionadas exclusivamente para tarefas de NLP são chamadas de *Transformer*. Dentro da terceira seção, há uma subseção na qual é tratado o uso do *Transformer* BERT, que é uma arquitetura de *Deep Learning* própria para NLP, e a sua variação, chamada de BERTSUM, desenvolvida para uma maior performance com sumarização de texto.

O bloco de estado da arte é composto apenas pela quarta seção, na qual são apresentadas as pesquisas mais recentes que utilizaram *Deep Learning*, mais especificamente a arquitetura BERT para sumarização automática de texto.

Para ilustrar o entendimento desenvolvido no Capítulo 1 e no Capítulo 2, foi elaborado o um Mapa Conceitual para mostrar o relacionamento entre as temáticas e os assuntos, conforme a Figura 2, abaixo:

FIGURA 2: Mapa Conceitual dos Fundamentos da Introdução e do Referencial Teórico



Fonte: elaborado pelo autor.

2.1 Pressupostos e fundamentos do Resumo Automático de Texto

A informação pode ser utilizada de diversas formas e ter inúmeras aplicações, o que torna a sua recuperação pouco eficaz. Para solucionar esse problema, foram criados vários métodos de recuperação da informação. Dentre esses métodos podemos citar a indexação, que classifica os documentos a partir de termos presentes ou não no documento, mas que estão intrinsecamente ligados ao armazenamento e à recuperação da informação (Lancaster, 2003). Outro método de recuperação da informação é o resumo, entendido como “uma representação sucinta, porém exata, do conteúdo de um documento” (Lancaster, 2003, p. 100). Entre as finalidades do resumo, pode-se citar que ele facilita a seleção, poupa tempo ao leitor, substitui a leitura de um item e auxilia na identificação de itens pertinentes.

O resumo pode ser chamado de Extrato quando é construído a partir da extração de frases do próprio documento, enquanto o resumo em si é uma transcrição direta do texto do autor (Lancaster, 2003). Os resumos podem ser classificados como indicativos, informativos ou críticos, sendo que um mesmo resumo pode incorporar elementos de indicativos e de informativos. Quando é feito por computador, o resumo recebe o nome de Sumarização Automática de Texto, ou *Automated Text Summarization* (ATS) (Ramezani; Feizi-Derakhshi, 2014).

Esse processo também pode ser chamado de Elaboração Automática de Resumos, para o qual Luhn (1958) criou um método e adotou alguns procedimentos, que hoje são executados por meio de *Machine Learning* e *Deep Learning*. De acordo com os procedimentos de Luhn, as frases que contêm os fatores de significância mais altos são selecionadas e impressas, na sequência em que ocorrem no texto, a fim de formar o “resumo”. É possível estabelecer um ponto de corte para controlar a quantidade de frases selecionadas. Isso pode se basear em um número fixo de frases, no número de frases necessárias para atingir certo percentual do texto total do documento ou em outros diversos métodos.

De acordo com Joshi *et al.* (2019), vários métodos tradicionais de sumarização de texto propostos na literatura são baseados principalmente em recursos de engenharia humana, entre eles combinação de recursos estatísticos e linguísticos, como frequência de termos, comprimento de sentença e posição ou palavras de sinalização e sintagma. Um exemplo é a abordagem para a análise da estrutura funcional do resumo do texto apresentado como parte de um método de representação semântica da informação de documentos científicos e técnicos, de Maeda (1981).

No entanto, independentemente do método, a sumarização automática é uma questão de seleção de frases, no sentido de escolher as frases que melhor representem o

conteúdo do texto presente e de organizar as frases selecionadas para otimizar a clareza do resumo (Lancaster, 2003).

Para Rush *et al.* (1971), um resumo pode ser produzido pela rejeição de sentenças do original que são irrelevantes para o resumo, por métodos de seleção e de rejeição de sentenças, nos quais incluem-se inferência contextual, referência de interseção, critérios de frequência e considerações de coerência. De acordo com (Lancaster, 2003, p. 316), “as expressões extraídas podem ser empregadas como termos de indexação, ser listadas para formar um tipo de resumo, ou usadas para ligar os termos de um vocabulário controlado”.

Nesse sentido, Moens e Dumortier (2000) desenvolveram um método baseado em gramáticas de texto para resumos em domínios ilimitados de assunto, no qual o sistema extrai frases e declarações relevantes para inclusão nos resumos. O objetivo desses autores era melhorar a navegação em um banco de dados de resumos de artigo para selecionar artigos de revistas relevantes.

Allan *et al.* (2001) avançaram nesse tema ao agregarem uma visão temporal. O foco da pesquisa desses autores foi o monitoramento das mudanças na cobertura de notícias ao longo do tempo, por meio da extração de uma única frase de cada evento dentro de um tópico de notícias, em que as histórias são apresentadas uma de cada vez e as suas frases devem ser classificadas antes que a próxima história possa ser considerada. A isso eles chamaram de Resumos Temporais de Notícias. Segundo (Lancaster, 2003, p. 322), “esse tipo de rastreamento automático do desenvolvimento de uma notícia ao longo do tempo foi denominado ‘rastreamento de eventos’”.

À análise de texto e de linguagem por meio computacional é dado o nome de Processamento de Linguagem Natural (PLN), ou *Natural Language Processing* (NLP), e sua base está na interdisciplinaridade de conceitos encontrados na ciência da computação, na ciência da informação, na linguística, na matemática, na inteligência artificial, na lógica e na psicologia (Chowdhury, 2003; Joshi, 1991). Uma característica notável da história do Processamento de Linguagem Natural (PLN) por computador é “quanto do que agora consideramos certo em termos de tópicos de interesse estava lá no início; tudo o que faltou aos pioneiros eram computadores” (Wilks, 2005, p. 2, tradução nossa), inclusive para Sumarização Automática de Texto.

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. 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 seu sentido original (Goularte *et al.*, 2014, 2019; Syed; Gaol; Matsuo, 2021; Tan; Kieuvongngam; Niu, 2020; YANG *et al*., 2013).

2.2 Caracterização do resumo automático de texto

O rápido aumento na taxa de geração de dados experimentado no início do Século XXI foi proporcionado, em grande parte, pelo crescimento da Internet, que possibilitou que as pessoas adquirissem e compartilhassem informações de diversas fontes e formatos. As mídias sociais fomentaram esse crescimento, pois permitiram aos usuários a criação de seu próprio conteúdo, aumentando o número e o tamanho dos documentos eletrônicos disponíveis na WEB. O aumento exponencial de dados e de informações leva a uma sobrecarga, tornando a localização de documentos e a recuperação de informações um desafio (Alami; Meknassi; En-Nahnahi, 2019; Hark; Karci, 2020; John; Premjith; Wilscy, 2017; Tayal; Raghuwanshi; Malik, 2017).

Nesse contexto, várias ferramentas de *Natural Process Language* (NLP) são necessárias para analisar essa massa de dados gerados (Alami; Meknassi; En-Nahnahi, 2019). Dentre essas ferramentas há o Resumo Automático de Texto, com o objetivo de "encontrar um subconjunto de dados que contenha as informações de todo o conjunto" (Lamsiyah *et al*., 2021b, p. 1, tradução nossa). As abordagens relativas ao conceito de sumarização de texto permitem várias aplicações, como:

QUADRO 1 – Principais características da sumarização de texto

a) redução do tempo de leitura;
b) facilitação do processo de seleção e de pesquisa de documentos;
c) melhora na eficácia da indexação;
d) redução de vieses encontrados em sumarizadores humanos;
e) aplicação em sistemas de perguntas/respostas;
f) maior capacidade de processamento de documentos;

Fonte: elaborado pelo autor.

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* é o processo de condensar o texto fonte em uma versão mais curta, preservando seu conteúdo de informação e seu 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).

A pesquisa de sumarização foi investigada pela comunidade de NLP por quase meio século (Christian; Agus; Suhartono, 2016), e quando utiliza recursos computacionais é chamada de *Automated Text Summarization* (ATS) (Ramezani; Feizi-Derakhshi, 2014). O ATS é uma área ativa de pesquisa com foco na condensação de um texto grande em um texto menor, retendo as informações relevantes (Tan; Kieuvongngam; Niu, 2020). Essa definição mantém o padrão de TS.

Para Joshi *et al.* (2019, p. 200, tradução nossa), *Automatic Text Summarization* "visa representar documentos de texto 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". O termo *Automated Document Summarization* é um método para "detectar as informações mais importantes no texto e, posteriormente, condensá-las para facilitar o uso pelos leitores" (Hark; Karci, 2020, p. 2, tradução nossa). Essa visão é corroborada por John, Premjith e Wilscy (2017), primeiro, por utilizarem o termo *Automated Document Summarization*, e, também, por definirem esse termo como "o processo de geração da versão compacta do documento/documentos enquanto mantém os aspectos importantes dos documentos/documentos de entrada" (John; Premjith; Wilscy, 2017, p. 385, tradução nossa). Segundo Mutlu, Sezer e Akcayol (2020, p. 2, tradução nossa), o domínio do processamento do texto "é uma experiência boa e altamente interpretável para redução de dimensão". A definição de ATS é praticamente a mesma de TS, porém, ATS foca em documentos de texto de grande tamanho, enquanto TS foca mais no método, na redução do texto.

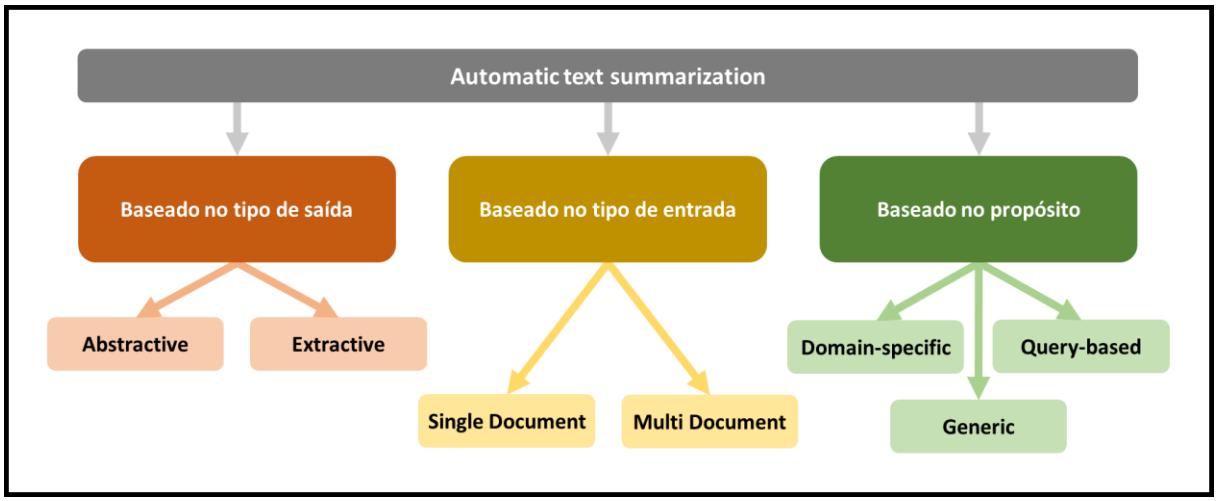
Existem várias técnicas de ATS e a maioria dessas abordagens modela o resumo de texto como um problema de classificação que resulta em incluir, ou não, a frase no resumo por métodos de pontuação baseados em recursos estatísticos, como TF-IDF (Frequência de termo-frequência inversa de documento) (Hark; 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 multidocumento não supervisionado extrativo a partir de sentenças salientes dos documentos de entrada, considerando a sumarização como um problema de otimização multicritério.

Segundo Lamsiyah *et al.* (2021b), a produção de resumos com maior diversidade de informações pode ser obtida ao melhorar a função de pontuação da frase combinando linearmente três métricas: a) relevância do conteúdo da frase; b) novidade da frase; c) e posição da frase. Já Protim Ghosh, Shahariar e Hossain Khan (2018), ao tratarem de resumos de notícias, introduziram recursos de pontuação de frase com base em gráfico. O resultado foram resumos com maior precisão e remoção de informações redundantes. Uma alternativa à pontuação da frase é utilizar Redes Neurais, que agrupam sentenças em um "espaço vetorial de alta dimensão para identificar grupos de sentenças semanticamente semelhantes entre si e selecionar representantes desses agrupamentos para formar um resumo" (Padmakumar; Saran, 2016, p. 1, tradução nossa).

O resumo de textos longos pode ser uma tarefa difícil e demorada quando executada por humanos, por isso a sumarização de texto é importante em vários campos, principalmente no empresarial e no acadêmico. Muitas funções da sumarização de texto estão relacionadas a várias tarefas de NLP, como Resposta a Perguntas, Classificação de Texto e outros. Segundo Sinha, Yadav e Gahlot (2018, p. 1, tradução nossa), "a geração de resumos é integrada a esses sistemas como uma etapa intermediária que ajuda a reduzir o comprimento do documento". A redução do texto torna a recuperação da informação mais rápida. Normalmente, resumos de textos são utilizados em motores ou em mecanismos de busca ou de pesquisa "para ajudar os usuários a escolher o conteúdo que melhor se adapta às suas necessidades de informação" (Alami *et al*., 2021, p. 196, tradução nossa). Esses resumos também são utilizados para gerar trechos como resultado da busca em um documento ou em *sites* de notícias, para facilitar a navegação, seja de processos judiciais, seja de resumos de textos biomédicos e clínicos (Syed; Gaol; Matsuo, 2021).

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 à extensão do resumo, que pode servir para criar um título ou um conjunto de palavras-chave e para gerar uma sequência curta, mas coerente de frases. As principais classificações referentes ao ATS foram sistematizadas na Figura 3, abaixo:

FIGURA 3: Tipos de métodos de *Automatic Text Summarization*



Fonte: elaborado pelo autor.

Com base no desenho da Figura 2, o ATS pode ser classificado em 3 tipos diferentes. Quando baseado no tipo de saída do resumo, classifica-se como *Extractive* ou *Abstractive*. Quando baseado na quantidade de documentos de entrada, classifica-se como *Single Document* ou *Multi Document*. Quando baseado no propósito, pode ser do tipo *Generic*, *Query-based* e *Domain-specific*. As seções seguintes detalham, de forma mais específica, cada tipo de ATS de acordo com sua classificação.

2.2.1 Método baseado no tipo de saída do resumo

A literatura aponta basicamente duas abordagens gerais de categorização do ATS quanto à saída. Para o termo Categorização, também são encontrados os termos de Métodos Básicos, de Classificação Ampla de Técnicas de Sumarização e de Classes. As duas formas de classificar as técnicas de sumarização de texto são realizadas a partir das seguintes categorias: a) *Abstractive summarization*; b) *Extractive summarization*. Ambas as formas são amplamente citadas e adotadas por vários pesquisadores de ATS (Christian; Agus; Suhartono, 2016; John; Premjith; Wilscy, 2017; Joshi *et al*., 2019; Padmakumar; Saran, 2016; Protim Ghosh; Shahariar; Hossain Khan, 2018; Sinha; Yadav; Gahlot, 2018; Syed; Gaol; Matsuo, 2021; Tan; Kieuvongngam; Niu, 2020; Yang *et al*., 2020).

A primeira abordagem, *Abstractive Summarization*, concentra-se na geração de novos resumos que parafraseiam o texto de origem, enquanto a segunda abordagem tem o objetivo de extrair e concatenar trechos importantes do texto de origem (Tan; Kieuvongngam; Niu, 2020). Entretanto, ambas as abordagens compartilham o propósito

comum de gerar resumos fluentes, não redundantes e coerentes, e podem ser empregadas para gerar resumos em um ou em vários documentos de origem (Syed; Gaol; Matsuo, 2021).

De forma mais específica, a abordagem *Abstractive Summarization* gera um resumo parafraseando o conteúdo principal do texto de origem em vez de escolher apenas parte do texto original, usando técnicas de geração de linguagem natural (Hark; Karci, 2020; Joshi *et al*., 2019; Tan; Kieuvongngam; Niu, 2020). O método *Abstractive Summarization* pode gerar frases sumárias, que não estão presentes no texto original (Padmakumar; Saran, 2016), ou até mesmo novas palavras, que não estão no texto original (Yang *et al*., 2020). Essa visão é corroborada por Song, Huang e Ruan (2019, p. 858, tradução nossa), pois esse método "gera sentenças escritas por humanos de maneira mais qualitativa para gerar resumos do zero sem estar restrito a frases do texto original".

O método *Abstractive Summarization* é mais desafiador, pois utiliza representação semântica para examinar e para interpretar o texto original e, em seguida, para usar essa representação semântica para gerar uma paráfrase, criando um resumo mais próximo do que um ser humano pode gerar (Christian; Agus; Suhartono, 2016; Protim Ghosh; Shahariar; Hossain Khan, 2018; Tan; Kieuvongngam; Niu, 2020). Esse método entende as sentenças de entrada e gera, por si só, as sentenças de resumo correspondentes (Iwasaki *et al*., 2020), tentando produzir um resumo ascendente com sequências de palavras que podem, ou não, estar presentes no texto original (Sinha; Yadav; Gahlot, 2018). Esse resumo pode tanto incluir inovações verbais, como fazer inferências a partir do texto de origem (Sinha; Yadav; Gahlot, 2018; Tan; Kieuvongngam; Niu, 2020).

Diante dessas características, o método *Abstractive Summarization* é mais complexo, pois, em primeiro lugar, envolve modelagem de linguagem complexa (Sinha; Yadav; Gahlot, 2018), por meio do uso de abordagem linguística, como cadeia lexical, rede de palavras, teoria dos grafos e agrupamento, para compreender o texto original e gerar o resumo (Joshi *et al*., 2019). Além disso, sua complexidade se dá pela falta de recursos de linguagem natural, já que essa abordagem "precisa de grandes quantidades de recursos linguísticos e ontologias geradas por humanos" (Alami; Meknassi; En-Nahnahi, 2019, p. 196, tradução nossa).

O método *Extractive Summarization* pode identificar e selecionar um subconjunto de palavras, de sentenças ou de frases relevantes do texto original e as encadear para formar um resumo, combinando algumas frases como estavam no texto original (Christian; Agus; Suhartono, 2016; Hark; Karci, 2020; Padmakumar; Saran, 2016; Protim Ghosh; Shahariar; Hossain Khan, 2018; Tan; Kieuvongngam; Niu, 2020; Yang *et al.*, 2020).

O método *Extractive Summarization* é chamado de Ordenação de Sentenças e consiste em ordenar e em extrair sentenças de acordo com as informações e os segmentos

mais importantes do texto original para combiná-los, formando um resumo coerente (Alami; Meknassi; En-Nahnahi, 2019; Song; Huang; Ruan, 2019). Esse método se baseia em uma função de pontuação (Sinha; Yadav; Gahlot, 2018) e mantém um grau razoável de gramaticalidade e de precisão (Tan; Kieuvongngam; Niu, 2020).

Segundo Christian, Agus e Suhartono (2016), o método *Extractive Summarization* é realizado por meio das seguintes abordagens estatísticas: a) método de título; b) método de localização; c) método de Frequência de termo-frequência inversa de documento (TF-IDF); e d) método de palavra para selecionar frases importantes ou palavras-chave do documento. De acordo com Joshi *et al.* (2019), ele geralmente consiste em três etapas principais: representação intermediária do texto de entrada, pontuação da frase e seleção da frase.

2.2.2 Método baseado no tipo de entrada do documento

Dependendo do número de documentos de origem a serem resumidos, os resumos podem ter como fonte um único documento (*Single Document*) ou vários documentos (*Multi Document*) (Hark; Karci, 2020; Lamsiyah *et al.*, 2021b). O resumo de documento único contém informações relevantes de um único documento, enquanto resumos de vários documentos contêm informações relevantes que cobrem todo o conceito de dois ou mais documentos sem muita redundância (John; Premjith; Wilscy, 2017).

2.2.3 Método baseado no propósito do resumo

Com base no propósito, os resumos podem ser classificados como T*ask-specific* (Q*uery-based*), focados na necessidade de um usuário com a consulta adaptada aos seus requisitos específicos ou de um grupo de usuários. A outra classificação é Genérica (G*eneric*), voltada a uma ampla comunidade de leitores (John; Premjith; Wilscy, 2017; Sarkar, 2009).

Há uma terceira opção de classificação em ATS: a *Domain-specific*, que se baseia no modelo D*omain-specific Knowledge* (Conhecimento Específico do Domínio) para formar um resumo com maior acurácia, como, por exemplo, ao resumir artigos de pesquisa de biomédicos a partir de termos e de frases específicas do domínio de conhecimento.

O termo é pouco utilizado na literatura, mas o conceito, não. O conceito de Conhecimento Específico do Domínio foi utilizado por Sinha, Yadav e Gahlot (2018), ao usarem modelos de resumos específicos de domínio ou de gênero, como relatórios médicos ou artigos de notícias, provando obter um resumo melhor. Esse é o mesmo caso encontrado

em Yang *et al.* (2020), ao observarem que os estilos de resumo para diferentes categorias de texto podem variar significativamente. Os autores exemplificam esse fato ao citarem que um resumo de política tende a enfatizar o assunto e o resultado ou a influência do evento, enquanto um resumo de esporte deve incluir as equipes e as pontuações do evento esportivo. Moradi, Dorffner e Samwald (2020), ao resumirem textos biomédicos, compararam um modelo genérico e um pré-treinado em texto biomédico, sendo que este obteve os melhores resultados.

Essa diferença acontece devido ao fato de as abordagens de sumarização de texto empregarem “um modelo uniforme para produzir resumos para os documentos fonte de diferentes categorias, que são propensos a gerar resumos genéricos e triviais que facilmente perdem ou sub-representam aspectos importantes dos documentos originais” (Yang *et al.*, 2020, p. 47, tradução nossa).

O método *Generic Summarization*, ou *Generic*, utiliza um número limitado de hipóteses para formar um resumo que tenta incluir o máximo de informação possível, salvaguardando o conteúdo geral do tema (Hark; Karci, 2020). O *Generic Summarization* fornece uma ideia geral do conteúdo do texto com base nos principais conceitos do documento, sem qualquer interação externa (John; Premjith; Wilscy, 2017; Joshi *et al.*, 2019), representando todos os fatos relevantes de um documento de origem sem considerar as necessidades de informação dos usuários (Lamsiyah *et al.*, 2021b).

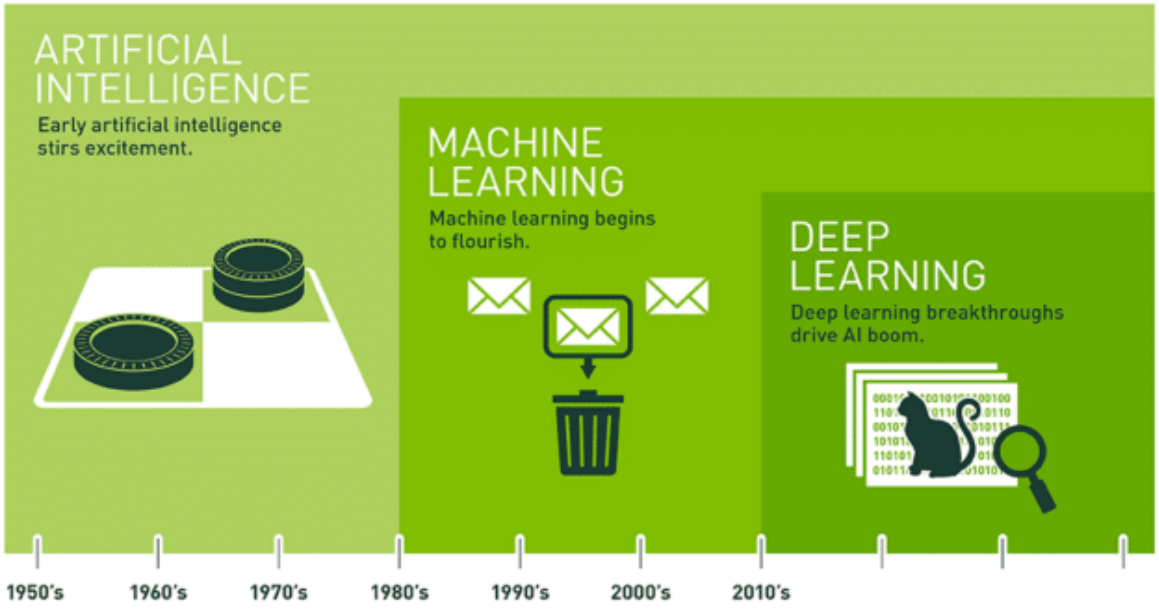
O outro método é o *Query-based Summarization*, derivado da necessidade de informações ou simplesmente da consulta do usuário (Lamsiyah *et al.*, 2021b). O *Query-based Summarization* gera resumos extraindo frases relevantes do documento de acordo com a consulta do usuário (John; Premjith; Wilscy, 2017; Joshi *et al.*, 2019), enfocando o tópico no qual ele está interessado e, portanto, recuperando informações do texto sobre o tópico relacionado (Hark; Karci, 2020).

2.3 Uso de *Deep Learning* para Resumo Automático de texto

A Inteligência Artificial oferece vantagens para lidar com problemas complexos associados a incertezas, para agilizar o processo de tomada de decisão, para diminuir as taxas de erro e para 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 se desenvolvido depois, e, finalmente, *Deep Learning,* contendo parte das duas (Copeland, 2016). Essa visão está representada na Figura 4, abaixo.

FIGURA 4: A Relação entre Inteligência Artificial, *Machine Learning* e *Deep Learning*



Fonte: Copeland, 2016.

O *Deep Learning* (DL) é um ramo, um tipo específico do *Machine Learning* e, 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 de Inteligência Artificial, pois o seu impacto foi sentido 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 suas principais características, conforme Quadro 2, abaixo:

QUADRO 2 – Principais características da evolução histórica do *Deep Learning*

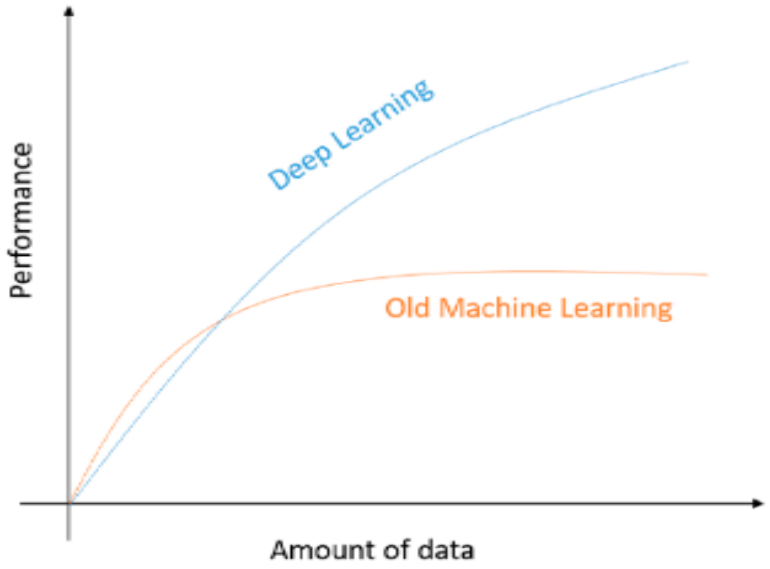
O *Deep Learning* tem uma longa e rica história, mas tem muitos nomes refletindo diferentes pontos de vista filosóficos, e aumentou e diminuiu em popularidade.
O *Deep Learning* se tornou mais útil conforme a quantidade e disponibilidade de modelos treinados aumentou.

O crescimento dos modelos de *Deep Learning* ao longo do tempo foi 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.

A principal distinção entre o ML tradicional e o DL está na forma como os recursos são extraídos, pois enquanto abordagens tradicionais de ML aplicam vários algoritmos de extração de recursos, em DL os recursos são aprendidos automaticamente e são representados hierarquicamente em vários níveis (Alom *et al*., 2019; Khamparia; Singh, 2019). Enquanto o aumento da quantidade de dados se torna um problema para algoritmos tradicionais de ML, por manter estável a performance, isso não acontece em DL, já que os algoritmos de DL aumentam o tamanho da rede e, por conseguinte, o seu desempenho (Alom *et al*., 2019; Khamparia; Singh, 2019). Essa ideia da diferença de performance está representada na Figura 5, abaixo:

FIGURA 5: A Performance de *Deep Learning* em Relação à Quantidade de Dados



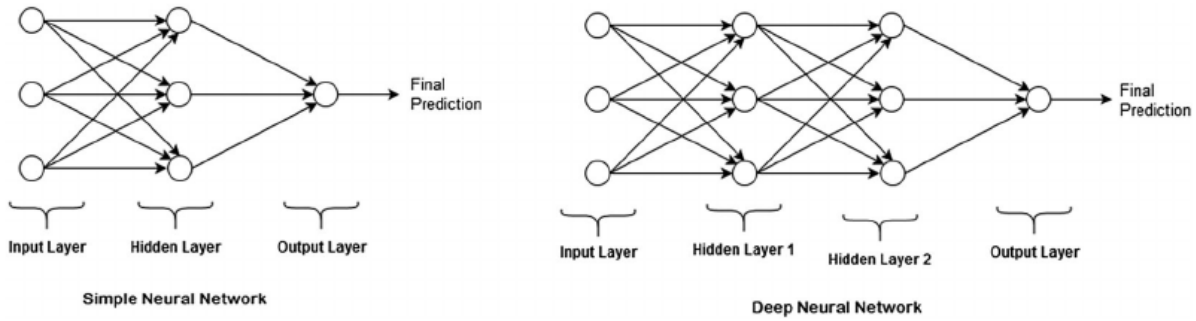
Fonte: Alom *et al.*, 2019, p. 7.

O ganho de performance à medida que os dados aumentam é possível por meio da utilização de modelos de rede neural de várias camadas (Leijnen; Veen, 2020), também denominados Arquiteturas de *Deep Learning*, que visam aprender a representação de recursos dos dados de entrada (Goodfellow; Bengio; Courville, 2016; Salehi; Burgueño, 2018). Uma arquitetura de *Deep Learning* utiliza uma entrada (*input*) para aprender a partir

de várias camadas de extração de recursos que processam as entradas, e esses recursos são computados por camadas profundas da rede, as quais são treinadas automaticamente por retropropagação para serem relevantes para a tarefa (Collobert *et al.*, 2011). 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 *Deep Neural Networks* (DNN) é um tipo de rede neural modelado como um *Perceptron* Multicamadas (MLP) treinado com algoritmos para aprender representações de conjuntos de dados sem qualquer projeto manual de extratores de recursos (Shrestha; Mahmood, 2019). A Figura 6 apresenta a principal diferença entre uma rede neural simples e uma *Deep Neural Networks* (DNN):

FIGURA 6: Diferença entre uma Rede Neural Simples e uma Rede Neural Profunda (DNN)



Fonte: Khamparia e Singh, 2019.

Essa modelagem de *Deep Neural Networks* (DNN) dá origem ao que é chamado de *Deep Learning*, um número maior de camadas de processamento, permitindo que funções mais complexas e não lineares sejam mapeadas. Consequentemente, as modernas arquiteturas de *Deep Learning*, que incluem 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).

O design dessas arquiteturas foi desenvolvido para suportar multicamadas, entre as quais existem milhares de interconexões e para utilizar técnicas de processamento de Unidades de Processamento Gráfico (*Graphics Processing Unit* – GPU), que capacita essas máquinas a executar arquiteturas profundas para detectar muitos recursos e suas interações (Khamparia; Singh, 2019).

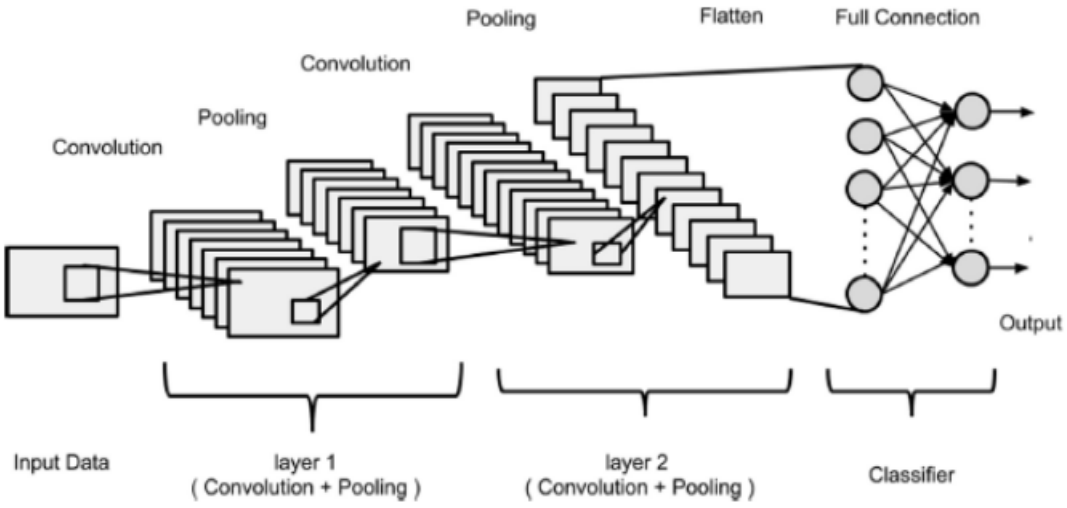
As modernas arquiteturas de *Deep Learning* implicam adicionar mais camadas e mais unidades dentro de uma camada, consistindo no mapeamento de um vetor de entrada

para um vetor de saída, criando modelos e conjuntos de dados suficientemente grandes com exemplos de treinamento rotulados (Goodfellow; Bengio; Courville, 2016). A partir dessa lógica de desenvolvimento das modernas arquiteturas de *Deep Learning*, inúmeras arquiteturas surgiram, com diferentes vantagens, com base na aplicação e nas características dos dados envolvidos (Sengupta *et al.*, 2020).

Os autores Khamparia e Singh (2019), Leijnen e Veen (2020) e Sengupta *et al.* (2020) relatam em seus trabalhos uma série de arquiteturas de Redes Neurais e de *Deep Learning*, como *Autoencoders*, *Convolutional Neural Networks* (CNN), *Recurrent Neural Networks* (RNN), *Hopfield Networks*, *Boltzmann Machines*, *Generative Adversarial Networks* (GAN), *Capsule Networks* e outras. Porém, algumas arquiteturas são preferidas em determinados casos, como, por exemplo, o uso de Redes Neurais Convolucionais (*Convolutional Neural Network* – CNN) para visão computacional e Redes Neurais Recorrentes (*Recurrent Neural Networks* – RNN) para sequência e para modelagem de séries temporais (Sengupta *et al*., 2020), sendo estas as mais modernas e as que mais se destacam (Goodfellow; Bengio; Courville, 2016; Zhang *et al*., 2020).

A arquitetura de *Deep Learning* chamada de Rede Neural Convolucional, ou *Convolutional Neural Network* (CNN), é composta de várias camadas, e o seu sucesso se deve a três propriedades importantes: a) campos receptivos locais; b) pesos compartilhados; e c) subamostragem espacial (Khamparia; Singh, 2019). Redes convolucionais contêm camadas convolucionais e de *pooling*, usadas para varredura aproximada de padrões, que são frequentemente correlacionados espacialmente (Van Veen *et al*., 2023). O descritivo do funcionamento da CNN é apresentado na Figura 7:

FIGURA 7: Camadas de Convolução e *Pooling* em CNN



Fonte: Sengupta *et al.*, 2020.

Existem várias arquiteturas de *Deep Learning* para a representação de texto relacionada 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 outras (Goodfellow; Bengio; Courville, 2016; Zhang *et al*., 2020).

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

2.3.1 Grandes Modelos de Linguagem – *Large Language Models* (LLMs)

A evolução das redes neurais aplicadas ao processamento de linguagem natural, também denominadas Modelos de Linguagem Neural, ou *Neural Language Models* (NLMs), teve sucesso devido à “representação vetorial dos *tokens*, que é aprendida automaticamente para otimizar a estimativa das probabilidades das palavras em uma sequência” (Guimarães; Campos; Jorge, 2024, p. 2, tradução nossa). Esse processo de aprendizagem relaciona palavras semanticamente relacionadas com vetores semelhantes, também chamados de *embeddings* de palavras (Guimarães; Campos; Jorge, 2024). O modelo Word2Vec é um exemplo inicial dessa arquitetura, e o avanço desse veio por meio da abordagem ELMo, que usa uma arquitetura RNN chamada de LSTM (*Long Short-term Memory*) “para aprender e produzir incorporações contextuais a partir do contexto dado pelas palavras à direita e das palavras à esquerda do *token*” (Guimarães; Campos; Jorge, 2024, p. 2, tradução nossa).

Após a implementação da arquitetura LSTM para NLP, o principal avanço ocorreu com os *Transformers*, ao utilizarem arquiteturas CNN, pois era permitido o uso de enormes *corpora* de treinamento, proporcionando alta eficiência em diversas tarefas de NLP. O principal exemplo desse avanço é o modelo BERT. Os *Transformers* que passaram a utilizar grandes quantidades de *corpus* para treinamento passaram a ser chamados de Grandes Modelos de Linguagem, ou *Large Language Models* (LLMs). Segundo Agathokleous *et al*. (2024, p. 210, tradução nossa), os LLMs são “baseados em parâmetros

abundantes, e esses modelos de *deep learning* são treinados com grandes quantidades de texto de forma não supervisionada".

De acordo com Shi *et al.* (2023), os LLMs construídos a partir de uma arquitetura de *Transformer* apresentam um aumento substancial em termos de parâmetros do modelo e dos dados de treinamento, permitindo-os capturar uma compreensão mais abrangente da linguagem. Para Huang, Wang e Yang (2023), a maioria dos textos contém algumas informações semânticas e sintáticas gerais, como gramática e significados de palavras comuns, e os LLMs podem ser pré-treinados em um grande *corpus* de texto não rotulado para aprender esses tipos de informações gerais, a partir de uma aprendizagem não supervisionada.

Segundo Shi *et al.* (2023), os LLMs possuem uma notável capacidade de aprendizado em poucos minutos e sem a necessidade de *fine-tuning*, podendo ser ainda mais alinhados com as preferências humanas mediante aprendizagem por reforço a partir do *feedback* humano, e esta abordagem foi implementada em vários LLMs, incluindo *Chat-GPT*.

Os GPT, *Generative Pre-trained Transformers*, são LLMs que surgiram em 2018 com o lançamento do GPT-1, porém, o grande avanço ocorreu no final de 2022, quando a OpenAI Inc. lançou o *Chat-GPT*, um *chatbot* que usa uma versão avançada do GPT-3 que é ajustada com aprendizagem supervisionada e com aprendizagem por reforço de *feedback* humano (Agathokleous *et al*., 2024). Os modelos GPT-3 e *Chat-GPT* "são treinados em grandes quantidades de dados de texto e alcançaram desempenho notável em uma variedade de tarefas de NLP, incluindo classificação de texto, resposta a perguntas e tradução automática" (Yang *et al*., 2023, p. 2, tradução nossa). De acordo com Shi *et al.* (2023), o "*Chat-GPT* enfatiza a adesão às instruções e à geração de respostas abrangentes, permitindo interações semelhantes às humanas altamente bem-sucedidas".

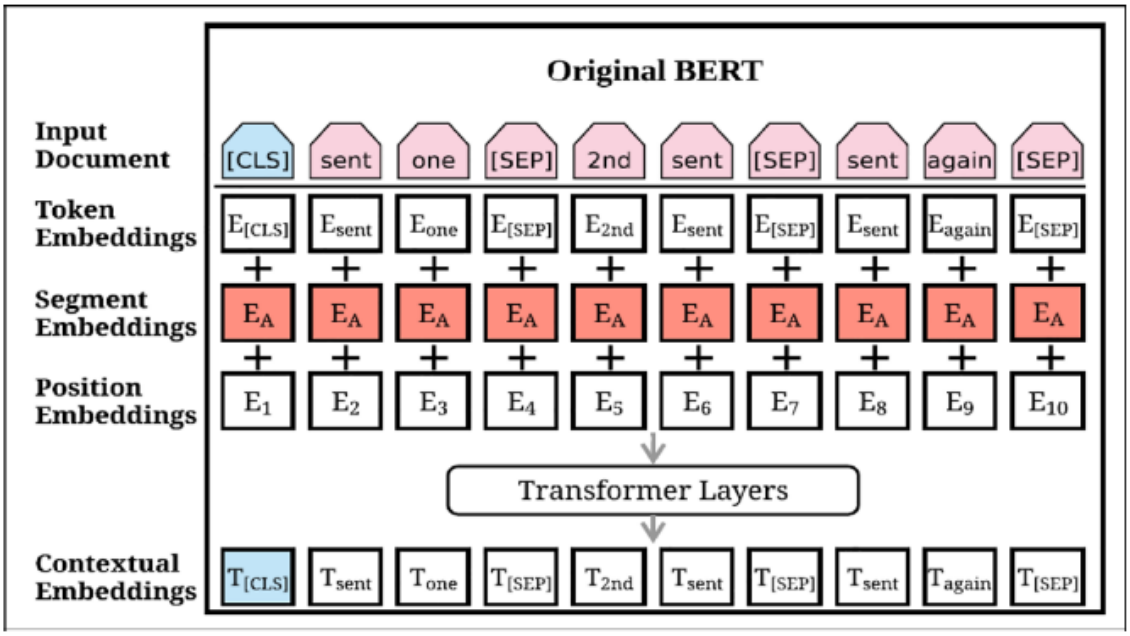
Por fim, os LLMs são basicamente algoritmos de *Deep Learning* que podem utilizar informações massivas para criar resultados inteligentes, como resumos, traduções e revisões (Agathokleous *et al*., 2024), e o desenvolvimento do GPT-4 ampliou significativamente os limites dos modelos de linguagem de última geração, pois mostrou o aprimoramento das habilidades de raciocínio, da compreensão de imagens e das capacidades multimodais, resultando em respostas mais sofisticadas e diversas (Shi *et al.*, 2023).

2.4 O *Transformer* BERT e a 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 novo, esse modelo consta em vários trabalhos de pesquisa, devido a sua versatilidade, tanto para processar muitos dados, quanto para permitir multi-idiomas e multi-tarefas de NLP.

O BERT é treinado em uma grande quantidade de dados de origem (cerca de 3,3 milhões de palavras) da *Wikipedia*, no idioma inglês, e no *BookCorpus*, usando duas tarefas não supervisionadas, incluindo a *Masked Language Modelling* e a previsão da próxima frase (Lamsiyah *et al*., 2021a). O BERT utiliza duas tarefas de pré-treinamento, sendo a primeira uma modelagem de linguagem mascarada, usada para criar representações em nível de *token*, e a segunda, uma previsão da próxima frase, para ensinar ao modelo dependências de longo prazo entre sentenças (Guimarães; Campos; Jorge, 2024). 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. A arquitetura original do modelo BERT é ilustrada na Figura 8.

FIGURA 8: Arquitetura Original do Modelo BERT



Fonte: Devlin *et al.,* (2018), Liu e Lapata (2019) e Lamsiyah *et al.,* (2021a).

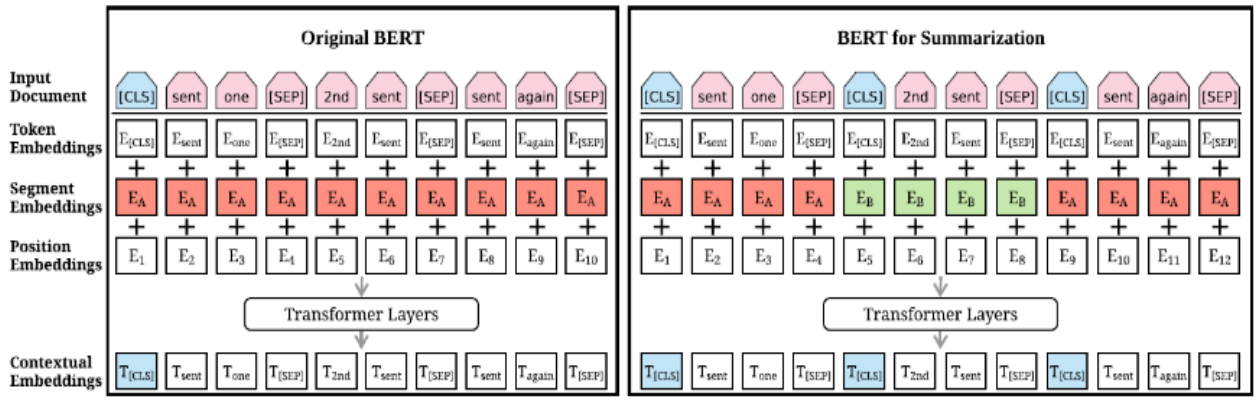
Dada a entrada S, que pode ser uma sequência de palavras, uma frase ou um par de frases reunidas, o codificador de léxico mapeia S em uma sequência de vetores de

embeddings, um para cada palavra, construída pela soma da palavra, pelos segmentos e pelos encaixes posicionais correspondentes. Em seguida, o codificador do *transformer* captura as informações contextuais de cada palavra por meio de *self-attention* e gera uma sequência de vetores de *embeddings* contextuais para a entrada S (Lamsiyah *et al*., 2021a).

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 por ter sido desenvolvida com objetivos específicos para o pré-treinamento. Em uma etapa, ele mascara aleatoriamente 10% a 15% das palavras nos dados de treinamento, tentando prever as palavras mascaradas, e na 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, entre os quais um tinha 110 milhões de parâmetros e o outro, 340 milhões (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, no qual 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. Por sua vez, na sumarização, deve-se codificar e manipular entradas multisentenciais. A Figura 9 ilustra a proposta de arquitetura BERT para SUMmarization (chamada pelos autores de BERTSUM).

FIGURA 9: Arquitetura Original BERT (esquerda) e BERTSUM (direita)



Fonte: Liu e Lapata, 2019, p.3.

O funcionamento do BERTSUM é baseado na representação de sentenças individuais, nas quais são inseridos *tokens* externos *[CLS]* no início de cada uma delas, e cada símbolo *[CLS]* coleta características para a sentença que o precede (Liu; Lapata, 2019). Também foram usadas incorporações de segmentos de intervalo para distinguir várias frases em um documento. Para $sent_i$ foram atribuídas a incorporação de segmentos E_A ou E_B dependendo se $_i$ é par ou ímpar. Por exemplo, para o documento [$sent_1$; $sent_2$; $sent_3$; $sent_4$; $sent_5$] é atribuído *embeddings* [E_A; E_B; E_A; E_B; E_A] e, dessa forma, as representações de documentos são aprendidas hierarquicamente onde as camadas de *Transformer* inferiores representam sentenças adjacentes, enquanto as camadas superiores, em combinação com a *self-attention*, representam o discurso multisentença (Liu; Lapata, 2019).

Os encaixes de posição no modelo BERT original têm um comprimento máximo de 512. Essa limitação foi superada, adicionando-se mais *embeddings* de posição, que são inicializados aleatoriamente e ajustados com outros parâmetros no codificador. O BERT adaptado denominado BERTSUM possibilita tanto *Extractive Summarization* quanto *Abstractive Summarization* (Liu; Lapata, 2019).

A sumarização extrativa pode ser definida como a tarefa de atribuir um rótulo y_i $\in$ *{0, 1}* para cada $sent_i$, indicando se a frase deve ser incluída no resumo. Assume-se que as frases resumidas representam o conteúdo mais importante do documento. Com BERTSUM, o vetor t_i, que é o vetor do símbolo $_{i\text{-}th}$ *[CLS]* da camada superior, pode ser usado como representação para $sent_i$. Várias camadas do *Transformer* entre frases são empilhadas em cima das saídas do BERT para capturar recursos de nível de documento para extrair resumos. A esse modelo foi dado o nome de BERTSUMEXT (Liu; Lapata, 2019).

Para *Abstractive Smmarization*, foi utilizada uma estrutura padrão de codificador-decodificador. O codificador é o BERTSUM pré-treinado e o decodificador é um *transformer* de 6 camadas, inicializado aleatoriamente. É concebível que haja uma incompatibilidade entre o codificador e o decodificador, uma vez que o primeiro é pré-treinado enquanto o segundo deve ser treinado sempre do início. Isso pode tornar o ajuste fino instável; por exemplo, o codificador pode superajustar os dados enquanto o decodificador não se adapta, ou vice-versa. Para contornar isso, foi projetado um novo escalonamento cronograma de *fine-tuning*, que separa os otimizadores do codificador e do decodificador (Liu; Lapata, 2019). Isso se baseia na suposição de que o codificador pré-treinado deve ser ajustado com uma taxa de aprendizado menor e com um decaimento mais suave, para que o codificador possa ser treinado com gradientes mais precisos quando o decodificador estiver se tornando estável.

Além disso, Liu e Lapata (2019) propuseram uma abordagem de ajuste fino em dois estágios, em que primeiro se ajusta ao codificador na tarefa de sumarização extrativa e

depois se ajusta a tarefa de sumarização abstrativa. Essa abordagem em dois estágios é conceitualmente muito simples, pois o modelo pode aproveitar as informações compartilhadas entre essas duas tarefas, sem alterar fundamentalmente sua arquitetura. O modelo padrão é chamado de BERTSUMABS e o modelo de dois estágios de ajuste fino, BERTSUMEXTABS.

2.5 Revisão do Estado da Arte em Resumo Automático de Texto

Segundo Vasconcellos, Nascimento da Silva e De 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, os *Transformers*, arquiteturas de *Deep Learning* desenvolvidas para processamento de linguagem natural, mudaram a forma como os modelos de linguagem são desenvolvidos.

Essas mudanças culminaram recentemente no lançamento de vários modelos de linguagem de grande porte (LLMs) ao público (Agathokleous *et al.*, 2024). Segundo Guimarães, Campos e Jorge (2024), os grandes modelos de linguagem (LLMs) impulsionaram substancialmente as pesquisas e as aplicações de Inteligência Artificial (IA) nos últimos anos e, atualmente, são capazes de alcançar alta eficácia em diferentes tarefas de processamento de linguagem natural (PNL), como tradução automática, reconhecimento de entidade nomeada, classificação de texto, resposta a perguntas ou resumo de texto.

No estudo dos autores Van Veen *et al.* (2023) foram aplicados métodos de adaptação a oito LLMs, abrangendo quatro tarefas distintas de resumo clínico, e esses resumos são equivalentes (45%) ou superiores (36%) em comparação com resumos de especialistas médicos. De acordo com os autores, “nossa pesquisa fornece evidências de que os LLMs superam os especialistas médicos na sumarização de textos clínicos em múltiplas tarefas” (Van Veen *et al.*, 2023, p. 1, tradução nossa).

Os autores Zhang, Liu e Zhang (2023) abordaram a limitação do resumo único devido ao fato de, por vezes, ser inadequada. Como solução, propuseram o *SummIt*, uma estrutura iterativa de resumo de texto baseada em grandes modelos de linguagem, como *Chat-GPT*. Essa estrutura permite que o modelo refine o resumo gerado de forma iterativa por meio de autoavaliação e de *feedback*, assemelhando-se ao processo iterativo humano ao redigir e revisar resumos.

Xie *et al.* (2023) focaram a pesquisa no Resumo de Texto Biomédico (BTS), desenvolvendo uma solução para apoiar a recuperação e a gestão de informação clínica. O BTS visa gerar resumos concisos que extraem informações importantes de documentos

biomédicos únicos ou múltiplos. Para os autores, essa pesquisa permitirá aos pesquisadores acompanhar rapidamente os avanços recentes e fornecer diretrizes para futuras pesquisas sobre BTS dentro da comunidade científica.

A pesquisa de Liu *et al.* (2023) investigou um novo cenário de aprendizagem de modelos de sumarização de texto que considera os LLMs como a referência, ou padrão ouro. Nos experimentos nos conjuntos de dados CNN/DailyMail e XSum, os autores demonstraram que modelos de resumo menores podem alcançar desempenho semelhante aos LLMs sob avaliação baseada em LLM. No entanto, os autores descobriram que os modelos menores ainda não conseguem atingir o desempenho do nível LLM sob avaliação humana, apesar das melhorias promissoras trazidas pelos métodos de treinamento propostos. Também foi realizada uma meta-análise sobre esse novo ambiente de aprendizagem, que revelou uma discrepância entre a avaliação humana e a avaliação baseada no LLM, destacando os benefícios e os riscos desse cenário de LLM como referência investigada.

Segundo os autores Yang *et al.* (2023), o surgimento de grandes modelos de linguagem (LLMs), como *GPT3* e *Chat-GPT*, criou um interesse significativo no uso desses modelos para tarefas de resumo de texto. No entanto, o desempenho dos LLMs para aplicações mais práticas, como resumos baseados em aspectos ou em consultas, é pouco explorado. Para preencher essa lacuna, os autores conduziram uma avaliação do desempenho do *Chat-GPT* em quatro conjuntos de dados de referência amplamente utilizados, abrangendo diversos resumos de postagens do *Reddit*, artigos de notícias, reuniões de diálogo e histórias. Nos experimentos, revelaram que o desempenho do *Chat-GPT* é comparável aos métodos tradicionais de ajuste fino em termos de pontuações do *Rouge*. Foram destacadas, ainda, algumas diferenças únicas entre resumos gerados pelo *Chat-GPT* e pelas referências humanas, fornecendo informações valiosas sobre a superioridade do *Chat-GPT* para diversas tarefas de resumo de texto.

Para os autores Wu *et al.* (2023), a maioria dos métodos de avaliação automática como *BLUE/ROUGE* pode não ser capaz de capturar adequadamente dimensões baseadas em vários critérios, tanto objetivos, como gramática e correção, quanto subjetivos, como informatividade, sucinta e apelo. Os autores propõem uma nova estrutura de avaliação baseada em LLMs, que fornece uma estrutura de avaliação abrangente, comparando o texto gerado e o texto de referência tanto em relação a aspectos objetivos como subjetivos. Os resultados experimentais em três conjuntos de dados reais para resumo mostraram que o modelo é altamente competitivo e tem uma consistência muito alta com anotadores humanos.

Segundo Yang *et al.* (2023), vários estudos investigaram o uso de grandes modelos de linguagem para tarefas de resumo de texto, e estudos recentes destacaram o potencial de linguagem de grandes modelos para sumarização de textos. Porém os autores destacam a necessidade de investigar melhor seu desempenho em diversas tarefas de sumarização em vários domínios. Diante disso, focou-se o trabalho em contribuir para esta pesquisa em andamento, avaliando as capacidades do *Chat-GPT* em tarefas de resumo baseadas em aspectos e em consultas e fornecendo *insights* sobre seus pontos fortes e suas limitações.

As atuais pesquisas com LLMs surgiram como um desenvolvimento notável no NLP, inspirando-se em modelos pré-treinados anteriores enquanto operam em escalas significativamente maiores e introduzem novos avanços (Shi *et al*., 2023). Entretanto, tomados em conjunto, estes resultados sugerem um grande potencial para os LLMs, que exigem a realização de mais pesquisas para examinar sistematicamente as características dos resumos gerados por LLMs, principalmente quando utilizado o *Chat-GPT* e suas variantes. Além disso, essas pesquisas ainda dependem de uma extensa avaliação humana, como sugerem Liu *et al.* (2023), Van Veen *et al.* (2023) e Zhang, Liu e Zhang (2023).

Com a popularização do uso das LLMs, especialmente quando focadas em resumo automático de texto, a tendência é o surgimento de cada vez mais artigos e pesquisas aplicando-as em um contexto especificado, propondo uma nova arquitetura ou a melhoria dos atuais LLMs. Essas pesquisas tendem a focar na obtenção de um indicador de qualidade do modelo melhor ou de maior aderência na resolução de um problema de alguma área do conhecimento.

Por serem relativamente recentes, as principais pesquisas de LLMs e de resumo automático de texto datam de 2022. Essas pesquisas ainda demandam algumas validações, principalmente quando aplicadas a contextos específicos. A expectativa é o surgimento de mais LLMs focadas em tarefas de NLP em áreas do conhecimento, como Medicina, Finanças, Jurídica e outras, e a vinculação de tarefa de NLP na área do conhecimento. Devido a esse contexto, as LLMs devem seguir o mesmo caminho do BERT, lançado em 2018, que se popularizou rapidamente por causa da sua versatilidade.

Devido à recente aplicação dos LLMs em pesquisas e ao fato de ainda não haver a maturação necessária para avaliar e para comparar qual modelo de LLMs e de resumo automático de texto seria o mais aderente a essa pesquisa, optou-se por desenvolver uma estratégia de busca concentrada no que havia de mais relevante em termos de Sumarização de Texto e de *Deep Learning* por meio do *Transformer* BERT.

Com base nessa busca, foi identificada a pesquisa de Abdel-Salam e Rafea (2022) um estudo sobre o desempenho de variantes de modelos baseados em BERT na

sumarização de texto a partir de uma série de experimentos e propuseram o “SqueezeBERTSum”, um modelo de sumarização treinado e ajustado com a variante codificadora SqueezeBERT, que alcançou pontuações competitivas do *ROUGE*, mantendo o desempenho do modelo de linha de base BERTSum em 98%, com 49% menos parâmetros treináveis. Dados os resultados dos experimentos na metodologia, existe uma versão potencializada do resumidor extrativo SqueezeBERT a partir dos resultados registrados acima. O SqueezeBERT tem menos parâmetros que o DistilBERT em aproximadamente 20% e produz a mesma pontuação *ROUGE-1* enquanto produz pontuações *ROUGE-2* e *ROUGE-L* ligeiramente mais altas. Embora o SqueezeBERT e o DistilBERT produzam pontuações ligeiramente mais baixas no modelo BERT-baseline, o SqueezeBERT tem a vantagem de ter menos tempo de treinamento e menos parâmetros do que o modelo de linha de base em ~48,44%, o que é quase metade dos parâmetros treinados do modelo BERT-baseline.

Bondielli e Marcelloni (2021) propuseram uma metodologia para representar o perfil de currículos profissionais de candidatos a emprego baseada em arquiteturas de sumarização e em *transformers* para geração de *embeddings* de currículos e em algoritmos de agrupamento hierárquico para agrupar esses *embeddings*. Os autores 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, eles estimaram a entropia da linguagem clínica usando *GPT-2*, que é um modelo anterior de linguagem causal autorregressivo 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.* (2021a), 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, os autores 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. Ajustou-se, também, o modelo BERT em tarefas intermediárias

supervisionadas mediante conjuntos de dados de *benchmark* GLUE usando métodos de ajuste fino de tarefa única e de multitarefa.

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* 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 em rede de memória dinâmica alcança um resultado comparável com outros sistemas extrativos nos conjuntos de dados.

Ma *et al.* (2020), ao discutirem sobre a dificuldade em tratar dependência de textos longos e em utilizar o mapeamento de tópicos latentes para modelos de sumarização de texto, desenvolveram um modelo específico para essa tarefa. Esses autores propuseram um modelo de sumarização extrativo e abstrativo baseado em tópicos chamado T-BERTSum, com base em Representações de Codificadores Bidirecionais de Transformadores (BERTs). Nesse modelo, a representação codificada do tópico latente, por meio do modelo de tópico neural (NTM), é combinada com a representação incorporada do BERT para guiar a geração com o tópico. Em segundo lugar, as dependências de longo prazo são aprendidas a partir da rede do *transformer* para explorar conjuntamente a inferência de tópicos e a sumarização de texto de uma maneira ponta a ponta. Em terceiro lugar, as camadas de rede de memória de longo prazo (LSTM) são empilhadas no modelo extrativo para capturar informações de tempo de sequência, e as informações efetivas são filtradas ainda mais no modelo abstrativo por meio de uma rede fechada. Além disso, um modelo extrativo abstrativo de dois estágios é construído para compartilhar as informações. Resultados experimentais nos conjuntos de dados CNN/Daily Mail e XSum demonstram que o modelo proposto alcança novos resultados de última geração enquanto gera tópicos consistentes em comparação com os métodos mais avançados.

Moradi, Dorffner e Samwald (2020) propuseram um método de sumarização utilizando o *transformer* BERT e concluíram que houve uma melhora significativa no desempenho do resumo de texto biomédico em comparação com um conjunto de métodos específicos e independentes de domínio. Para isso, eles combinaram diferentes versões do BERT com um método de agrupamento para identificar as sentenças mais relevantes e informativas dos documentos de entrada.

Moradi, Dashti e Samwald (2020) abordaram os desafios da sumarização de textos baseada em gráficos no contexto biomédico utilizando diferentes tipos de *embeddings* contextualizados e livres de contexto, como Word2vec e GloVe, e um tipo específico de

BERT, o BioBERT. O BioBERT é um modelo de linguagem desenvolvido por meio de um pré-treinamento de BERT com base em grandes *corpora* de texto biomédico para mapear o texto de entrada para *embeddings* contextualizados. O *transformer* BERT possui variações com foco no idioma, na tarefa de NLP e na área do conhecimento.

You, Zhao e Chen (2020) propuseram uma abordagem de sumarização de alta qualidade que se concentra no conteúdo do tópico do documento e na semelhança entre o resumo e o documento de origem, com foco em sumarização abstrativa. Para tal propósito, foi utilizado o BERT, primeiro, para extrair palavras-chaves de tópicos e as fundir com documentos de origem como parte da entrada e, em segundo lugar, para aproximar o resumo do documento fonte, calculando a semelhança semântica entre o resumo gerado e o documento fonte, melhorando a qualidade do resumo. Os autores conseguiram maximizar a pontuação de similaridade por meio de treinamento, obtendo bons resultados no conjunto de dados LCSTS e gerando um resumo mais legível e coerente.

Srikanth *et al.* (2020) utilizaram o modelo BERT para produzir sumarização extrativa, agrupando os *embeddings* de sentenças por agrupamento *K-means*. Dessa forma, os autores introduziram um método dinâmico para decidir o número adequado de sentenças para escolher os agrupamentos. O objetivo desses autores visou produzir resumos de maior qualidade, incorporando resolução de referência e produzindo resumos dinamicamente de tamanhos adequados, dependendo do texto.

A busca na literatura acadêmica retornou poucos trabalhos que refletem o tema dessa pesquisa. As pesquisas apresentadas acima mostram o intensivo uso de BERT para a sumarização de texto, porém, a maioria se refere à criação de uma metodologia de sumarização ou ao melhoramento do BERT para se tornar mais performática nessa tarefa. 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 da elaboração de uma metodologia de recuperação da informação a partir da aplicação de sumarização automática de texto para monitorar a evolução de mudanças estruturais no macroambiente de um setor. Pode-se concluir que, apesar de já existir uma quantidade significativa de pesquisa sobre *Deep Learning* aplicando o *transformer* BERT, a aplicação para monitorar mudanças estruturais em um setor ao longo do tempo ainda não foi amplamente explorada.

3 METODOLOGIA

O presente 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, segundo Lakatos e Marconi (2003, p. 83), “o método é o conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo - conhecimentos válidos e verdadeiros - traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista”.

Quanto ao método, a pesquisa se classifica como indutiva, por partir “de dados ou observações particulares constatadas, podendo chegar a proposições gerais” (Richardson, 2012, p. 35). Segundo Gil (2008, p. 10), nesse método, “parte-se da observação de fatos ou fenômenos cujas causas se deseja conhecer” e “procura-se compará-los com a finalidade de descobrir as relações existentes entre eles”.

Quanto à finalidade/natureza, a pesquisa se classifica como aplicada, pois o objetivo é “gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos” (Prodanov; Freitas, 2013, p. 51).

Quanto aos objetivos, a pesquisa se classifica como exploratória, já que visa “examinar 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 qualiquantitativo, por usar, concomitantemente, métodos quantitativos e qualitativos (Creswell, 2014).

A classificação da pesquisa permite direcionar o esforço para enfrentar o desafio de lidar com massivas quantidades de dados textuais em grandes conjuntos de dados digitais, “visando recuperar em tempo hábil informações relevantes para algum objetivo específico” (Souza, 2005, p. 3). Para isso, 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 a notícias referentes ao setor de Mineração de Minerais no Brasil.

Enquanto escopo de pesquisa, esse trabalho abordará o uso de notícias nos idiomas português e inglês e não adotará o uso de fluxo dinâmico de entrada e de tratamento de notícias. Será tratado um bloco estático de notícias.

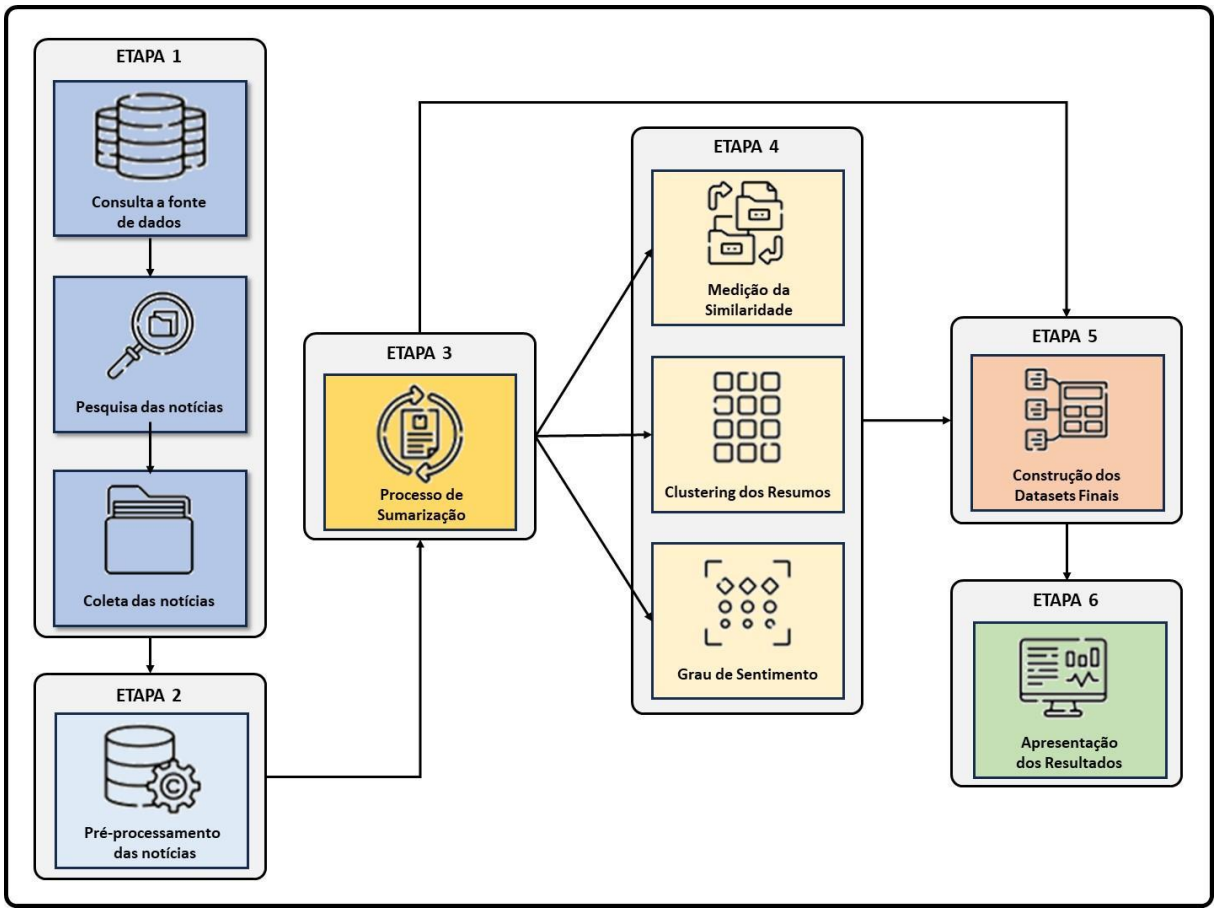
A metodologia será dividida em duas etapas, sendo a primeira de construção do modelo e a segunda de validação, junto a especialistas.

Sendo assim, esse capítulo é composto por duas seções. Na primeira seção, são apresentadas as etapas metodológicas da construção do processo de sumarização das notícias e de aplicação de ferramentas subsequentes. Na segunda seção, são apresentadas as etapas metodológicas de validação do modelo construído na seção anterior, inclusive a elaboração da ferramenta de visualização de dados e de validação do modelo junto a especialistas do Setor de Mineração.

3.1 Descrição das Etapas da Construção do Modelo

Este trabalho apresenta um método de recuperação da informação a partir da aplicação de sumarização automática de texto para monitorar a evolução de mudanças estruturais no macroambiente do setor de mineração no Brasil. Como se trata de uma pesquisa classificada como método misto, a Figura 10, abaixo, descreve o procedimento de extração e os métodos quantitativos aplicados aos dados coletados. Na sequência, será estruturado um painel de dados para agrupar os dados coletados que, então, será submetido à avaliação de especialistas do setor de Mineração.

FIGURA 10: Etapas da Construção do Modelo



Fonte: elaborado pelo autor.

O *Transformer* BERT foi aplicado nas tarefas de Grau de Similaridade Textual e de Classificação de Sentimento, e na tarefa de Sumarização de Texto foi aplicado o BERTSUM, que é uma variação do BERT, parametrizado para a sumarização de texto. A técnica de *Machine Learning* de Clusterização por meio do algoritmo *K-means* foi adotada para agrupar as notícias e extrair as 10 principais por *Cluster*. A Figura 10 fornece uma visão geral da metodologia adotada nessa fase de construção do modelo, composta por 6 etapas, 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 terceira, 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.

3.1.1 Pesquisa e Coleta das notícias

A primeira etapa consiste em extrair um *corpus* de dados composto por notícias variadas sobre mineração, sendo que cada notícia é um documento, para a realização do processamento estabelecido. As notícias-alvo do setor de mineração do Brasil serão obtidas de uma base de notícias de uma empresa proprietária de informações. Essa empresa proprietária monitora mais de 10 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 à abrangência de países, de setores e de fontes, essa empresa será escolhida.

Nessa fonte de dados serão pesquisadas notícias do setor de mineração no Brasil, excluindo termos como *criptomoedas, bitcoin, data mining*, mineração de processos e suas respectivas combinações. A *query* de consulta utilizada foi a seguinte: (mineração, or *mining*) not (criptomoedas or *cripto* or *bitcoin* or "*data mining*" or "mineração de dados" or "*process mining*" or "mineração de processos"). Cada notícia retornada dessa pesquisa será exportada para um arquivo de texto.

3.1.2 Pré-processamento das notícias

Cada arquivo de notícia terá a sua estrutura adequada, de forma que todas as reportagens sigam 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

íntegra. Esse processo permitirá retirar notícias com apenas data e título. O passo seguinte será transformar cada notícia em um arquivo em PDF.

A etapa seguinte será a construção de um fluxo de importação de partes de cada notícia em PDF para um arquivo tabular. Nesse arquivo tabular serão imputados a data, a fonte, o título e o tamanho da notícia em quantidade de caracteres. A notícia na íntegra não será imputada para não aumentar o tamanho do arquivo tabular. O campo “data” da notícia passará por um processo de validação para ter o formato data em cada notícia. Ao final, ter-se-á uma tabela com os seguintes campos: Referência do arquivo PDF, Data, Fonte, Título da Notícia e Tamanho da Notícia.

3.1.3 Processo de Sumarização

A linguagem de programação escolhida foi o *Python*, que foi desenvolvido no começo dos anos 90 e, desde então, teve várias modificações. Atualmente é uma das principais ferramentas utilizadas para o desenvolvimento de ciências de dados, seja por engenheiros, por cientistas ou por pesquisadores. O *Python* permite a inserção de bibliotecas, também chamadas de Pacotes, que possuem rotinas computacionais específicas para determinadas tarefas, como, por exemplo, tratamento de dados tabulares, por meio da biblioteca Pandas.

Para processar os resumos, a tabela com os dados das notícias será particionada em arquivos de, no máximo, 250.000 caracteres totais de notícias.

O processo de sumarização será construído para chamar cada partição da tabela e, de cada partição, extrair apenas a notícia do arquivo PDF, procedendo a sumarização.

Para a sumarização, será utilizado o BERTSUM, uma variação do BERT desenvolvida especificamente para proceder à sumarização de texto. Em termos de bibliotecas do *Python*, para a execução da tarefa de sumarização será utilizada a *“Summarizer”* para proceder a Sumarização Extrativa. Para cada referência de arquivo PDF na tabela de dados será imputada uma sumarização da notícia original, e, na sequência, os arquivos particionados serão integrados, formando um novo arquivo. Finalmente, será inserido um novo campo contendo o tamanho da sumarização em termos de quantidade de caracteres.

3.1.4 Tarefas secundárias de NLP

Com a tabela finalizada com os resumos das notícias, 3 processos de NLP serão aplicados para se obter os objetivos específicos. Os processos de NLP serão: a) Medição da Similaridade textual; b) Classificação do Grau de Sentimento da notícia; e c) *Clustering* dos Resumos.

A Similaridade Textual é utilizada para determinar o quão semelhante dois textos são. Os modelos de similaridade textual com BERT convertem os textos de entrada em vetores (*embeddings*) que capturam informações semânticas e calculam o quão próximos (semelhantes) eles estão. A saída é uma pontuação de similaridade entre 0 e 1, sendo que quanto mais próximo de 1, mais semelhantes os textos são. Para essa tarefa será construído um fluxo no qual a primeira notícia da tabela será comparada com as demais, gerando uma pontuação de similaridade para cada par de notícias, e isso ocorrerá para cada uma delas. A pontuação de similaridade de cada par de notícias será classificada observando a seguinte regra descrita no Quadro 3:

QUADRO 3 – 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
0,000 a 0,099	Nula

Fonte: elaborado pelo autor.

A partir dessa classificação dos pares de notícias, será contada a quantidade de notícias por tipo de similaridade. Se uma notícia obtiver similaridade igual a zero, significa que ela não é similar a nenhuma outra com uma pontuação entre 0,95 e 1,00.

O BERT também pode ser utilizado para classificar um texto quanto ao seu sentimento, que pode ser positivo, negativo ou neutro. O BERT mensura o sentimento do texto gerando uma pontuação entre 1 e 5. Esse método foi aplicado a cada sumarização, de forma que cada notícia recebeu uma pontuação entre 1 e 5. Essa pontuação foi classificada para transformar uma variável contínua em uma discreta, sendo que valores 1 e 2 foram classificados como negativo, 3 como neutro e 4 e 5 como positivo.

O último processo aplicado será 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* será escolhida a técnica de *K-means*. Em algoritmos de *Clustering*, 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 1, o *cluster* é denso e bem separado dos outros. 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 a *clusters* errados. Cada *cluster* será nomeado de acordo com as suas características para possibilitar uma comparação ao longo do tempo.

3.1.5 Construção dos *Datasets* para análise

Ao final, serão obtidos dois arquivos de dados, sendo um composto por uma chave única para cada notícia, pela sumarização dessa notícia, pelo seu grau de sentimento e pela quantidade de notícias por tipo de similaridade. O segundo arquivo também terá uma chave única para notícia, de forma que os arquivos possam ser relacionados, além dos resultados do *Clustering*.

3.1.6 Apresentação dos Resultados

A partir das duas tabelas**,** será possível construir gráficos e tabelas para analisar os resultados, como também criar um modelo de visualização dos dados para demonstrar a utilização da metodologia.

3.2 Fase de Validação do Modelo

Após a etapa de construção e análise do modelo, será realizada uma etapa de validação junto a especialistas do setor de Mineração e de *Analytics & Data Science*. O foco dessa etapa é validar o quanto o modelo proposto é capaz de responder à pergunta central desta pesquisa.

Essa etapa de validação se caracteriza como uma pesquisa qualitativa, por haver estratégias de investigação mais específicas, que “se concentram na coleta, na análise e na redação dos dados” (Creswell, 2014, p. 210).

A técnica de pesquisa escolhida foi a Observação Direta Intensiva, no formato de Entrevista. A entrevista “é um procedimento utilizado na investigação social, para a coleta de dados ou para ajudar no diagnóstico ou no tratamento de um problema social” (Lakatos; Marconi, 2003, p. 195). Segundo Gil (2008, p. 109), a entrevista enquanto técnica de coleta de dados “é bastante adequada para a obtenção de informações acerca do que as pessoas sabem, crêem, esperam, sentem ou desejam, pretendem fazer, fazem ou fizeram, bem

como acerca das suas explicações ou razões a respeito das coisas precedentes". De acordo com Lakatos e Marconi (2003, p. 196), "a entrevista tem como objetivo principal a obtenção de informações do entrevistado, sobre determinado assunto ou problema".

Em relação ao tipo de entrevista, foi selecionada a não-estruturada focalizada individual, caracterizada por proporcionar ao entrevistado "uma forma de poder explorar mais amplamente uma questão" (Lakatos; Marconi, 2003, p. 197), por conter perguntas abertas, que podem ser respondidas em uma conversação informal, seja face a face ou por telefone, e por haver um roteiro focado no problema que o entrevistador quer estudar. Segundo Creswell (2014), esses tipos de entrevista podem ser chamados de Entrevistas Qualitativas, e "envolvem questões não estruturadas e em geral abertas, que são em pequeno número e se destinam a suscitar concepções e opiniões dos participantes" (Creswell, 2014, p. 214). De acordo com Gil (2008, p. 114), "essas entrevistas são muito utilizadas em estudos exploratórios, com o propósito de proporcionar melhor compreensão do problema, gerar hipóteses e fornecer elementos para a construção de instrumentos de coleta de dados" e também "podem ser utilizadas para investigar um tema em profundidade, como ocorre nas pesquisas designadas como qualitativas". Richardson corrobora essa visão da seguinte forma:

> A entrevista não estruturada, também chamada de entrevista em profundidade, em vez de responder à pergunta por meio de diversas alternativas pré-formuladas, visa obter do entrevistado o que ele considera os aspectos mais relevantes de determinado problema: as suas descrições de uma situação em estudo. Por meio de uma conversação guiada, pretende-se obter informações detalhadas que possam ser utilizadas em uma análise qualitativa. A entrevista não estruturada procura saber que, como e por que algo ocorre, em lugar de determinar a frequência de certas ocorrências, nas quais o pesquisador acredita. (Richardson, 2012, p. 208)

O tipo de amostragem dessa pesquisa é por conveniência, que se caracteriza pelo fato de o pesquisador selecionar intencionalmente os indivíduos, os participantes que comporão o estudo, por entender que esses elementos o ajudarão a entender melhor a questão de pesquisa e por ter acesso a eles (Creswell, 2014; Gil, 2008).

Para as entrevistas, foi construído um roteiro de perguntas, conforme Quadro 4, abaixo, baseado no *dashboard* desenvolvido a partir das análises da etapa Construção do Modelo. O *dashboard* pode ser acessado pelo link https://luanderfalcao.wixsite.com/tese-de-doutorado.

QUADRO 4 – Roteiro de Perguntas

1 – Área: ( ) Mineração ( ) Dados e Informação
2 – E-mail:
3 – Idade (em anos):
4 – A quanto tempo você trabalha nesta área? (tempo total em anos)
5 – Qual o seu cargo atual? ( ) pleno ( ) sênior ( ) especialista ( ) coordenador ( ) supervisor ( ) gerente ( ) diretor
6 – Quais fontes de informação utiliza para acompanhar ou se informar quanto as questões setoriais? ( ) Jornais (exemplo: Folha de São Paulo, O Globo, Valor Econômico e semelhantes) ( ) Sites de notícias ( ) Newsletters ( ) Colegas ou amigos ( ) Relatórios/informativos internos ( ) Relatórios de consultorias ( ) Outras fontes
7 – Quais temas você tem o costume de acompanhar ou monitorar relacionados a setores produtivos: ( ) Novas tecnologias ( ) Mudanças na legislação ou normas regulatórias ( ) Desempenho da economia ( ) Questões ligadas ao meio ambiente e sociais ( ) Aspectos culturais ( ) Mudança de estratégia ( ) Parcerias, fusões e aquisições ( ) Outros
8 – A partir do uso do Dashboard, o quanto você acha que método de tratamento e apresentação da informação foi útil para ajudar no entendimento do setor?
9 – O método utilizado no Dashboard possui estrutura capaz de mudar o seu entendimento ou a sua percepção sobre o setor, ou de alguma parte do setor?
10 – Assinale quais temáticas foram percebidas/identificadas ao manipular o Dashboard: ( ) Novas tecnologias ( ) Mudanças na legislação ou normas regulatórias ( ) Desempenho da economia ( ) Questões ligadas ao meio ambiente e sociais ( ) Aspectos culturais ( ) Mudança de estratégia ( ) Parcerias, fusões e aquisições ( ) Outros
11 – As 10 principais palavras por Cluster por ano e por contexto retrataram os principais fatos ocorridos na mineração naquele ano, segundo a sua percepção?

Fonte: elaborado pelo autor.

Dadas as características dessa etapa da pesquisa, foram escolhidos propositalmente cerca de 4 especialistas em Mineração e 11 em Dados e Informação, com ênfase em *Analytics & Data Science*. Os especialistas em *Analytics & Data Science*

compreendem um conjunto de profissionais que atuam com vertentes de *Bussiness Intelligence (BI), Data Science, Machine Learning, Big Data, Artificial Intelligence e Analytics*. Esses especialistas foram selecionados pelo conhecimento e pelo domínio em suas respectivas áreas e por terem uma visão crítica capaz de avaliar a metodologia e dizer o quanto é aderente para responder à pergunta central da pesquisa.

Para cada especialista foi enviado um convite para a entrevista e o *link* do *dashboard*. Na sequência, era feito um agendamento para uma conversa. Na conversa, as perguntas de 1 a 7 eram realizadas para introduzir o assunto, e as perguntas de 8 a 11 eram feitas para induzir o entrevistado a falar, expor a sua experiência ao usar o *Dashboard*, e, consequentemente, abordar os fatos positivos e negativos identificados e o quanto o *Dashboard*, que materializa a metodologia, é capaz de responder à pergunta central.

Durante a entrevista, as respostas às questões 8, 9 e 11 foram agrupadas em Muito, Médio, Pouco e Não retratou, de acordo com a avaliação e a descrição do entrevistado. A fala do entrevistado quanto à questão 10 foi agrupada de acordo com as alternativas descritas no roteiro.

4 DISCUSSÃO DOS RESULTADOS DO MODELO CONSTRUÍDO

Ao todo, foram coletadas e inseridas em uma tabela 3.824 notícias, com datas entre 11/02/2003 e 30/11/2021, conforme descrito na metodologia. Foram retiradas as notícias com 0 caractere e com o título idêntico, permanecendo, ao final, 3.271 notícias. Ao agregar a quantidade de notícias por ano, optou-se por retirar aquelas que se encontravam entre os anos de 2003 e 2012, dada a baixa quantidade de notícias nesse período. Ao fim dessa etapa, obteve-se 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, 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 para 2,2 milhões de caracteres, com as notícias tendo redução entre 9% e 99%. Esse fato corrobora o fato de o ATS ser uma etapa anterior e importante no processamento textual, por reduzir o texto e possibilitar a utilização de outras tarefas, nesse caso, de Classificação e de Similaridade. Nos quadros 5 e 6, são demonstrados exemplos da sumarização por meio do BERTSUM, sendo apresentada a notícia na íntegra e a sua sumarização.

QUADRO 5 – Aplicação do BERTSUM em uma notícia – Exemplo 1

Data:	17/04/2020
Fonte:	Money Times – News
Título:	Vale produz 59,6 milhões de toneladas de minério de ferro no 1º tri e reduz projeção para 2020
Notícia na integra:	A produção de minério de ferro da Vale (VALE3) no primeiro trimestre somou 59,6 milhões de toneladas, queda de 18% ante o mesmo período do ano passado, ficando abaixo do *guidance* da empresa de 63-68 milhões de toneladas, informou a mineradora, que revisou ainda as metas de produção de seus principais minerais para o ano. O volume menor que o esperado para os primeiros três meses do ano, segundo a companhia, foi devido a diversas causas operacionais, além de condições climáticas mais severas e concentradas do que o habitual, principalmente em março. A empresa prevê agora produzir de 310 milhões a 330 milhões de toneladas de minério de ferro em 2020, ante projeção anterior de 340-355 milhões de toneladas.
Resumo:	A produção de minério de ferro da Vale (VALE3) no primeiro trimestre somou 59,6 milhões de toneladas, queda de 18% ante o mesmo período do ano passado, ficando abaixo do *guidance* da empresa de 63-68 milhões de toneladas, informou a mineradora, que revisou ainda as metas de produção de seus principais minerais para o ano. O volume menor que o esperado para os primeiros três meses do ano, segundo a companhia, foi devido a diversas causas operacionais, além de condições climáticas mais severas e concentradas do que o habitual, principalmente em março.

Fonte: elaborado pelo autor.

QUADRO 6 – Aplicação do BERTSUM em uma notícia – Exemplo 2

Data:	17/06/2020
Fonte:	CMA Agency - News
Título:	MINERAÇÃO:PPI qualifica dois projetos de mineração como prioridade nacional
Notícia na integra:	São Paulo, 17 de junho de 2020 – O Programa de Parcerias de Investimentos (PPI) qualificou os projetos de mineração Leilão Cobre de Bom Jardim e Fosfato de Miriri como empreendimentos de prioridade nacional e colocou os empreendimentos em consulta pública. Segundo o órgão, os projetos localizados em Goiás, Pernambuco e Paraíba, serão colocados a leilão para cessão de direitos minerários. Hoje eles são títulos ativos do Serviço Geológico do Brasil (CPRM), empresa pública vinculada ao Ministério de Minas e Energia (MME). As audiências públicas serão realizadas de forma virtual e os documentos ficarão em consulta pública até o próximo dia 29 de junho. Wilian Miron / Agência CMA
Resumo:	São Paulo, 17 de junho de 2020 – O Programa de Parcerias de Investimentos (PPI) qualificou os projetos de mineração Leilão Cobre de Bom Jardim e Fosfato de Miriri como empreendimentos de prioridade nacional e colocou os empreendimentos em consulta pública. Segundo o órgão, os projetos localizados em Goiás, Pernambuco e Paraíba, serão colocados a leilão para cessão de direitos minerários.

Fonte: elaborado pelo autor.

Para cada resumo, foi mensurada a similaridade semântica de uma notícia com as demais. Para essa etapa foi construída uma relação de pareamento das notícias, obtendo-se 10.390.952 pares, ou seja, uma notícia foi relacionada com as demais 3.223. Para processar esse volume de dados de 10 milhões de pares, foram criados 208 arquivos de 50.000 pares, a partir do particionamento da base de pares. Esse particionamento foi necessário para proceder o processamento, pois a estimativa para mensurar a similaridade dos 50.000 pares de cada arquivo foi de 18 minutos, totalizando 60 horas de processamento. Para esse processamento, foi utilizada uma máquina de RAM 25Gb, Disco 166 Gb e GPU T4 RAM Alta.

Cada mensuração de similaridade foi inserida no arquivo com o seu respectivo par de notícia. Ao final, os 208 arquivos foram agrupados, transformando cada valor de similaridade em uma variável categórica, conforme Quadro 3. Na etapa seguinte, foi contada a quantidade de notícias por Tipo de Similaridade, criando-se uma tabela com o nome do arquivo em PDF. Essa nova estrutura de dados foi incorporada à tabela com os resumos das notícias.

Foi considerado que quando uma notícia tiver Tipo de Similaridade Perfeita igual a 0 (zero), isso significa que ela não possui similaridade com nenhuma outra. Ao retirar as notícias com alto teor de similaridade, obtêm-se 2.315 notícias. As notícias com alto teor de similaridade, 909 ao todo, indica serem notícias muito parecidas, logo, não sendo necessária a leitura ou o acesso dessas para recuperar a informação.

Cada resumo de notícia da tabela foi classificado quanto ao grau de sentimento. Foi utilizado o *transformer* BERT de classificação de sentimento, que gera uma pontuação entre 1 e 5. O resumo de notícias com pontuação 1 e 2 foi classificado como negativo, pontuação 3 como neutro, 4 e 5 como positivo. Também foi realizado um comparativo de tempo de processamento dessa tarefa entre a notícia original e o seu respectivo resumo. Utilizando um *notebook* Colab do Google, com um processador de dois núcleos, 12 *GBytes* de memória RAM e Disco de 166 *GBytes*, as 3.224 notícias originais consumiram 63 minutos de processamento, enquanto os resumos das notícias consumiram 19 minutos. Ao utilizar o mesmo recurso computacional para a mesma tarefa de NLP, o ATS executou em 1/3 do tempo. Isso corrobora o fato de ser o ATS uma etapa inicial importante por reduzir o texto original, gerando ganho de tempo de processamento sem perda informacional.

Por fim, a estrutura da tabela de resumos permitiu recuperar a informação dos resumos por data, por tipo de similaridade, por grau de sentimento e por redução da notícia.

Também foi criada uma tabela a partir do *Clustering* das notícias por contexto. A estrutura da tabela de resumos permite criar contextos, ou seja, segmentos de análise baseados em certas características da base de dados, como grau de similaridade e data. A partir do grau de similaridade e data, foram gerados 6 contextos, sendo: a) Geral; b) Geral por ano; c) Geral por mês; d) Sem similaridade; e) Sem Similaridade por ano; e f) Sem Similaridade por mês.

Em cada contexto, as notícias foram agrupadas em *clusters* e receberam o valor do *Cluster*, um valor inteiro variando entre 1 e 5, e cada *Cluster* recebeu um nome de acordo com as suas características. Com o propósito de gerar um gráfico de dispersão dos *Clusters* por contexto, cada notícia recebeu um valor de eixo X e Y a partir do cálculo de *Principal Component Analysis* (PCA).

Foram extraídas, ainda, as 10 palavras mais representativas de cada *Cluster* por contexto. Essas palavras indicam qual assunto aquele *Cluster* de notícias trata. O objetivo é identificar se os *Clusters* tratam do mesmo assunto, ou não, ou se possuem, ou não, um discurso semelhante.

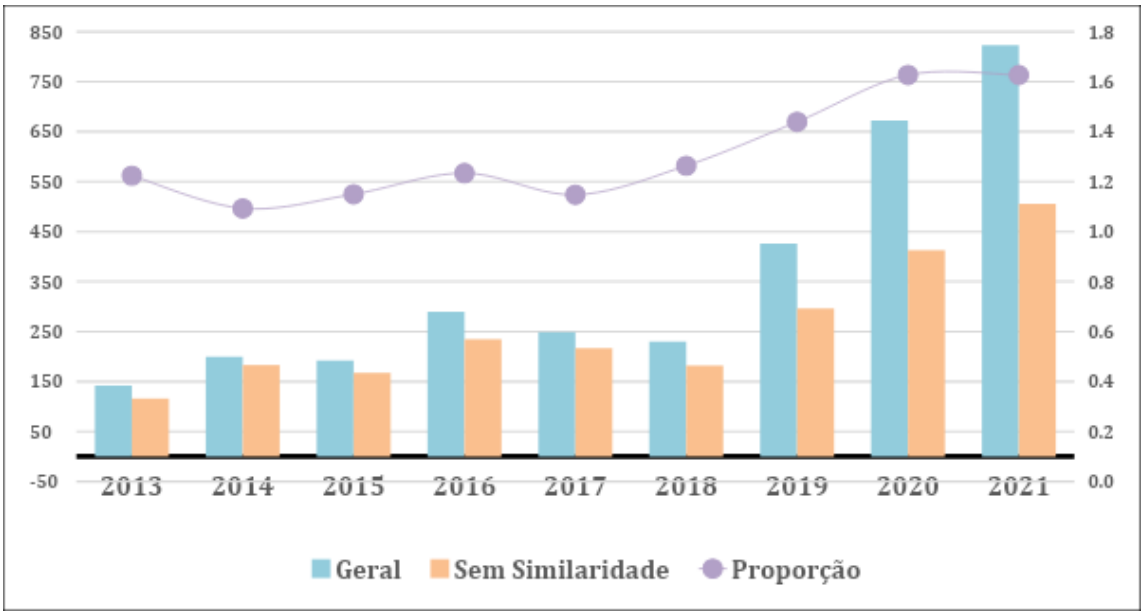
4.1. Exploração dos Resultados Iniciais

Ao analisar a evolução da quantidade de notícias por ano, tanto no contexto Geral, com 3.224 notícias quanto no contexto Sem Similaridade, com 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, refletindo 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 de notícias do contexto Geral pela quantidade de notícias do contexto Sem Similaridade por ano. O aumento dessa proporção indica uma maior repetição semântica das notícias. Esses resultados podem ser observados no Gráfico 1, abaixo:

GRÁFICO 1 – Evolução da Quantidade de Notícias por Ano, Contexto Geral e Sem Similaridade

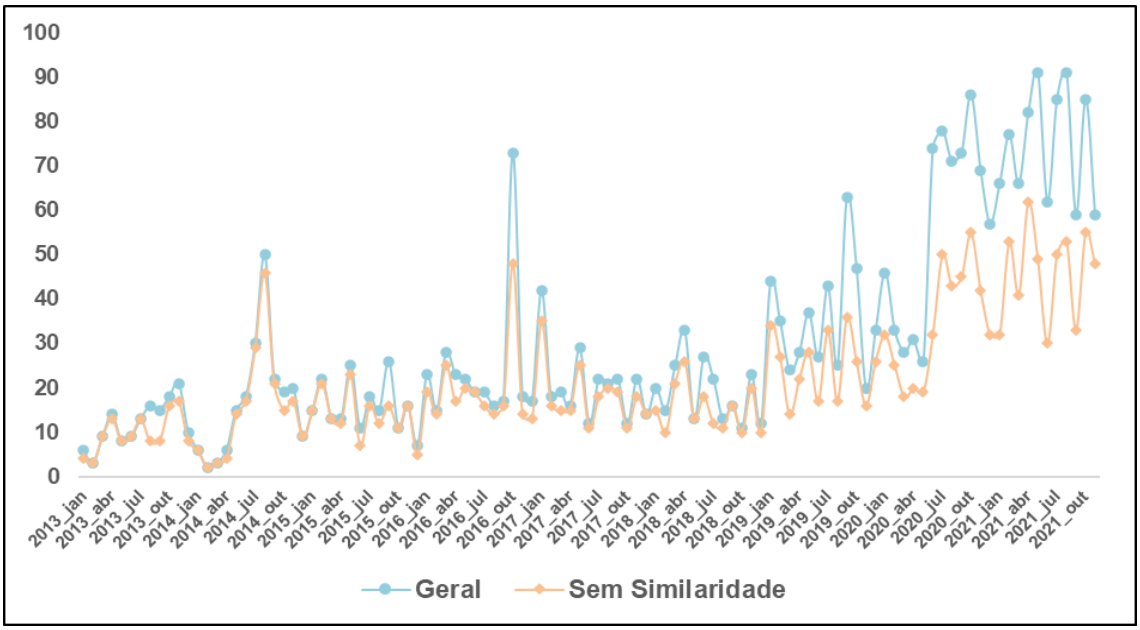


Fonte: elaborado pelo autor.

A percepção do aumento da repetição das notícias no mesmo período fica mais evidente quando se analisa a quantidade de notícias por mês, no contexto Geral e no contexto Sem Similaridade, conforme Gráfico 2, abaixo. No gráfico, a quantidade de notícias do contexto Geral passa a se descolar da quantidade de notícias do contexto Sem Similaridade a partir de 2019, o que não ocorria nos períodos anteriores.

Uma das justificativas para esse fato é o aumento do uso das mídias sociais, que criam um volume maior de informações e de notícias e as disseminam rapidamente. Além disso, para se criar uma notícia e disseminá-la, não é mais necessário ser um jornalista, o que torna qualquer pessoa com certo conhecimento do tema um potencial gerador de informação. Além disso, quando a notícia é importante e possui relevância, ela tende a ser mais repetida, aumenta a velocidade de sua disseminação.

GRÁFICO 2 – Evolução da Quantidade de Notícias por Mês, Contexto Geral e Sem Similaridade



Fonte: elaborado pelo autor.

Toda essa conjuntura endossa a dedução de uma maior repetição semântica do conteúdo das notícias a partir do ano de 2019. Em ambos os gráficos de evolução da quantidade de notícias os dados mostram uma influência do tempo ou de fatores ligados ao período que deve ser considerado quanto à análise.

Ao analisar a classificação do Sentimento das notícias, a maioria das notícias foram classificadas como negativas, sendo 63,90% no contexto Geral e 60,60% no contexto Sem Similaridade. Os dados mostram uma redução da quantidade de notícias classificadas como negativas do contexto Geral para o Sem Similaridade, e, consequentemente, um aumento das notícias classificadas como positivas e neutras. Esse resultado indica uma disseminação maior de notícias com teor semântico negativo, corroborando a necessidade de uma análise de similaridade semântica para evitar uma análise baseada apenas na quantidade de ocorrências.

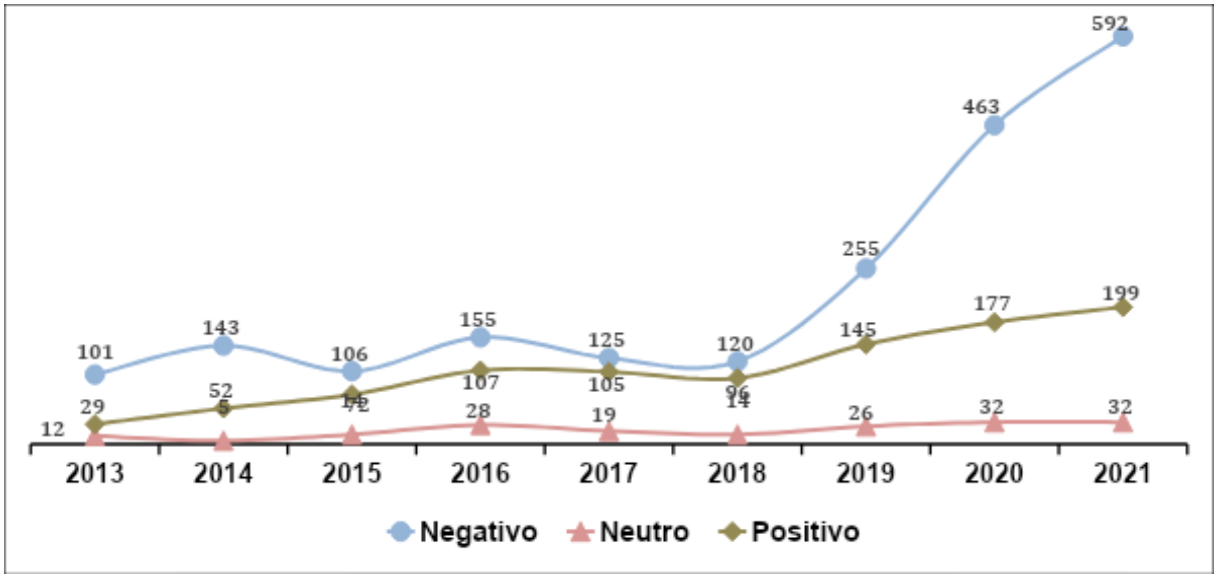
Tabela 1 – Quantidade de Notícias por Grau de Sentimento por Contexto Geral e Sem Similaridade – em %

Sentimento	Contexto	
	Geral	Sem Similaridade
Negativo	63,90%	60,60%
Neutro	5,65%	6,26%
Positivo	30,46%	33,13%

Fonte: elaborado pelo autor.

Uma outra forma de analisar a classificação das notícias por grau de sentimento é distribuindo-as no tempo. No Gráfico 3, os dados do contexto Geral foram plotados, e as notícias classificadas como neutras se mantiveram estáveis desde 2016. As notícias classificadas como positivas crescem de forma mais consistente a partir de 2019. No gráfico, é possível evidenciar que, a partir de 2019, ano que começa a ter um volume maior de notícias, há uma proporção de duas notícias negativas para cada notícia positiva. No ano de 2020, essa proporção é de 2,6, e no ano de 2021, é de 3. Os dados mostram um aumento significativo de publicações de notícias classificadas como negativas, que são as mais disseminadas, por terem mais relevância, e, consequentemente, por serem mais repetidas naquela época específica.

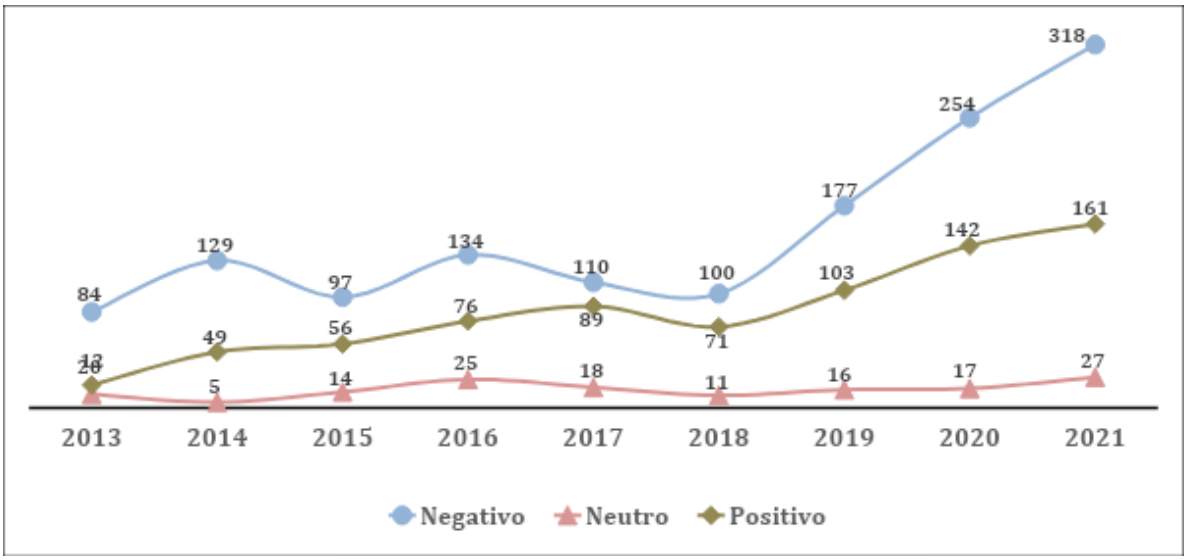
GRÁFICO 3 – Evolução da Quantidade de Notícias por Grau de Sentimento Contexto Geral



Fonte: elaborado pelo autor.

No contexto Sem Similaridade, no qual notícias redundantes foram retiradas e foram mantidas apenas aquelas com complementariedade, os dados do Gráfico 4 mostram que as notícias classificadas como neutras mantiveram um patamar de estabilidade ao longo dos anos. As notícias classificadas como positivas crescem de forma mais consistente a partir de 2018. No gráfico, é possível evidenciar que, a partir de 2019, ano em que começa a haver um volume maior de notícias, há uma proporção de 1,7 notícias negativas para cada notícia positiva. No ano de 2020, essa proporção é de 1,8 e no ano de 2021, essa proporção é de 2. Esses resultados podem ser observados no gráfico.

GRÁFICO 4 – Evolução da Quantidade de Notícias por Grau de Sentimento Contexto Sem Similaridade

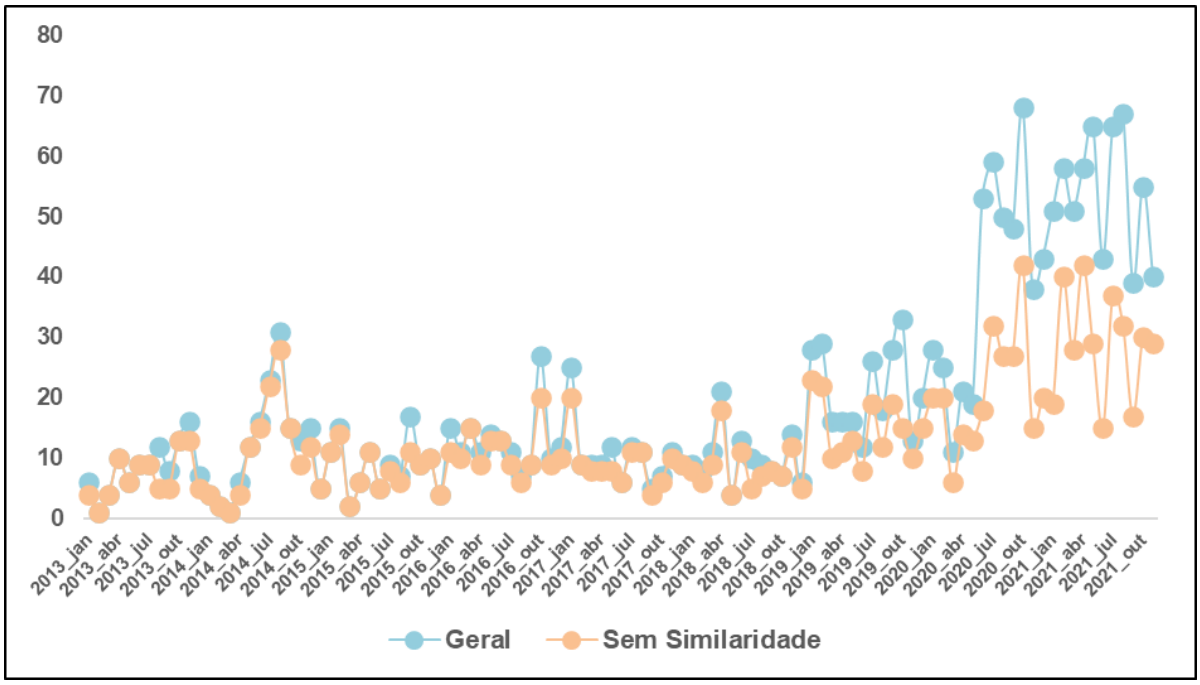


Fonte: elaborado pelo autor.

Ao comparar apenas as notícias positivas dos gráficos 3 e 4, os dados mostram um crescimento maior entre 2019/2020 e 2020/2021 do contexto Sem Similaridade em relação ao contexto Geral. Ao analisar a taxa de notícias negativas por notícias positivas, no contexto Geral há uma taxa maior do que no contexto Sem Similaridade, especialmente quando se analisam apenas os anos de 2019, de 2020 e de 2021. Esse fato evidência que a replicação de notícias negativas tende a mascarar, ou até mesmo influenciar na análise de grau de sentimento quando se utiliza apenas esse parâmetro para análise de notícias. Logo, percebe-se a necessidade de retirar a redundância e fazer um refinamento da informação por meio do contexto Sem Similaridade.

Os dados também mostram que a repetição das notícias é mais acentuada entre aquelas classificadas como negativas. Ao analisar a quantidade de notícias por mês do contexto Geral e Sem Similaridade, o resultado é um afastamento entre os dois contextos, principalmente a partir de 2019, conforme o Gráfico 5:

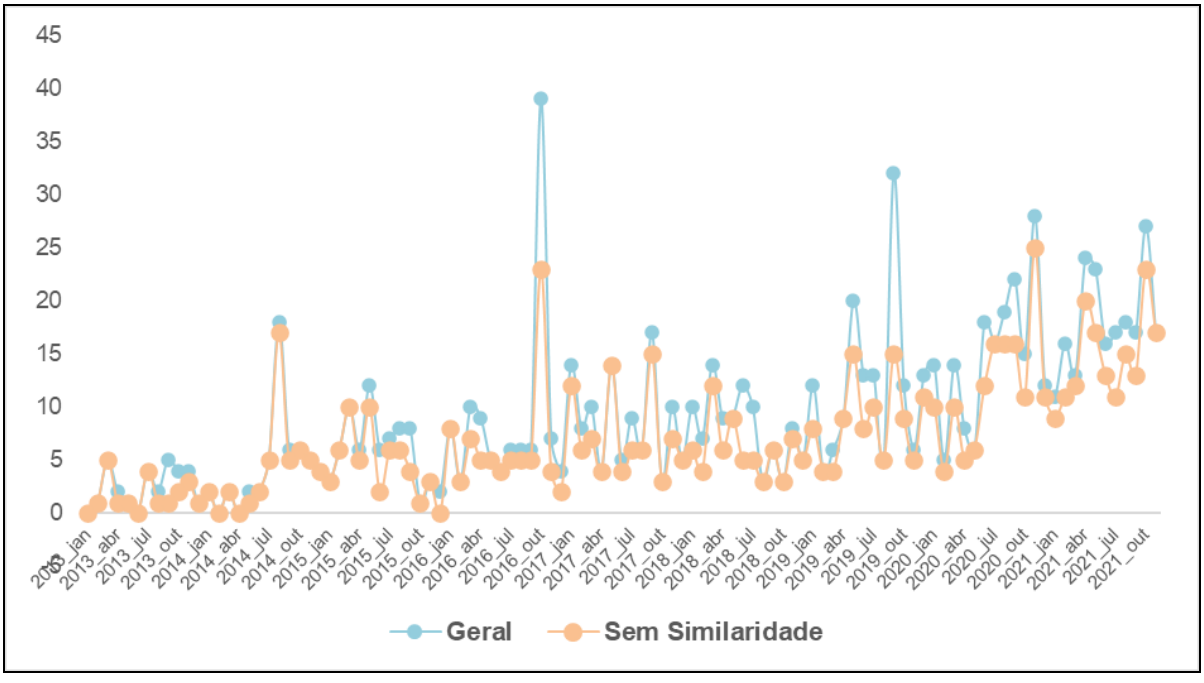
GRÁFICO 5 – Evolução da Quantidade de Notícias Classificadas como Negativas por Mês, e por Contexto Geral e Sem Similaridade



Fonte: elaborado pelo autor.

Ao proceder a mesma análise para as notícias classificadas como positivas, os dados mostram uma evolução mais padronizada, mesmo com picos em certos períodos, conforme Gráfico 6, abaixo:

GRÁFICO 6 – Evolução da Quantidade de Notícias Classificadas como Positivas por Mês, e por Contexto Geral e Sem Similaridade



Fonte: elaborado pelo autor.

Ao compararmos os gráficos 5 e 6, os dados mostram que há uma repetição semântica maior das notícias classificadas como Negativas do que das Positivas, fato esse que corrobora a influência da repetição das notícias classificadas como negativas na análise do todo. Esse fato leva à dedução de que há pouca produção de conteúdo de notícias, havendo apenas uma replicação por diversas fontes. Aliado a isso, há o crescimento das plataformas digitais, que ajudam na replicação dos conteúdos, e quanto mais replicados, maior indicação de interesse e de importância é dada àquela notícia ou conteúdo. Essa conjuntura justifica a necessidade de utilização de uma estratégia de sumarização de texto para analisar material textual.

Para entender qual a história ou qual o discurso presente nessa quantidade massiva de texto, foi utilizado o algoritmo *K-means* de *Clustering*, da categoria de *Machine Learning*, denominada *Unsupersived Learning*, para agrupar os resumos por padrões de texto, com o objetivo de extrair a principal mensagem. Além de agrupar as notícias, também foi possível extrair as 10 palavras mais representativas de cada *Cluster*.

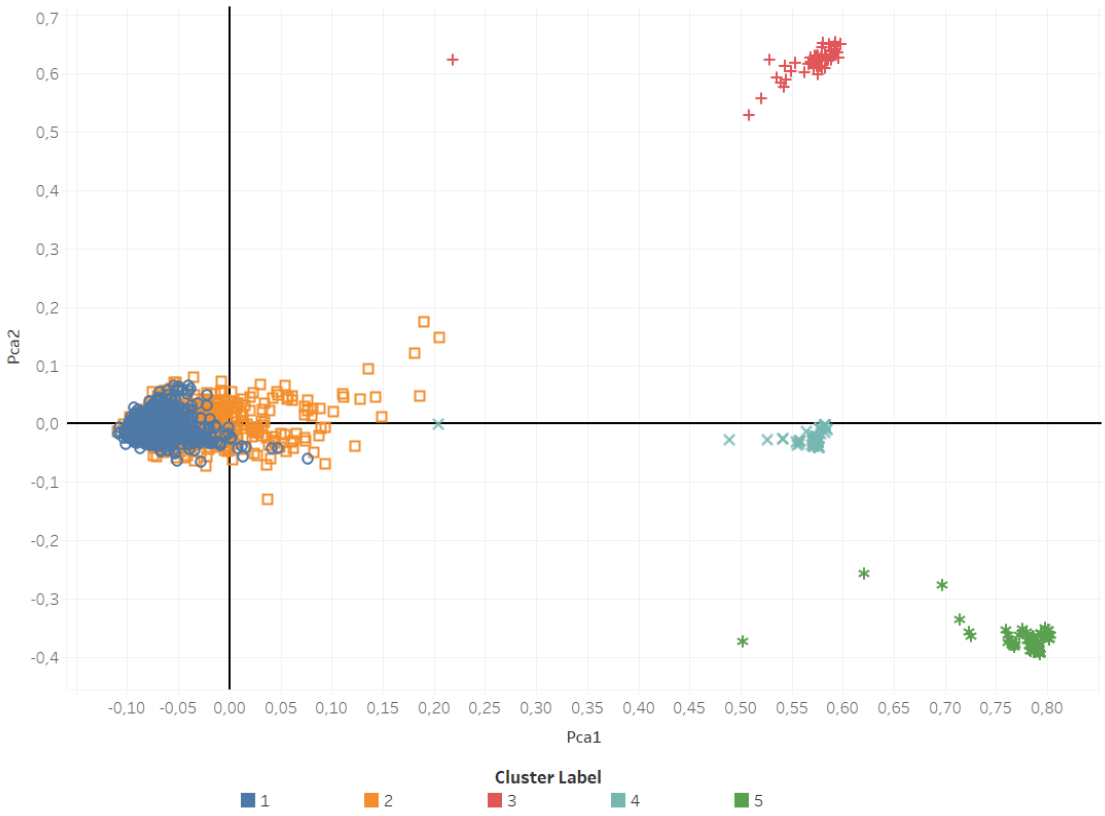
4.2 Apresentação e Análise dos Resultados no Contexto Geral e Sem Similaridade

Na construção dos *Clusters,* foram criados dois eixos, X e Y, por meio da Análise de Componentes, com valores representativos para cada notícia, conforme Gráfico 7. A pontuação *Silhouette* foi de 0,0552, evidenciando uma forte sobreposição entre os *Clusters*. Os dados do Gráfico 7 mostram uma sobreposição entre as notícias do *Cluster* 1 e 2, sendo esses os mais representativos, por congregarem cerca de 90% das notícias. Os demais *Clusters* estão mais distantes dos *Cluster* 1 e 2, possuem menos notícias e não estão sobrepostos. Esse resultado indica que os *Clusters* 1 e 2 tendem a contar a mesma história, ou a seguirem o mesmo discurso. No caso dos *Cluster* 3, 4 e 5, eles tendem a contar uma história diferente. De forma geral, pode-se deduzir que o agrupamento realizado apresenta um padrão na apresentação das notícias sobre o Setor de Mineração do Brasil.

O *Cluster 1* é caracterizado pela quantidade de notícias, cerca de 2.157 ou 66,90% do total. Ao analisar as palavras mais representativas desse *Cluster*, as notícias tratam do setor de mineração, no Brasil e em Minas Gerais, da empresa Vale e de um projeto que envolve minas. Pode-se inferir serem essas as principais palavras que ocorrem com maior frequência em uma notícia sobre mineração no Brasil, por abordar o estado de Minas Gerais e por esse ser o principal estado produtor de minério de ferro, e a empresa Vale, por ser a maior mineradora do país. Em termos de quantidade de notícias por ano,

cerca de 57% delas estão concentradas nos anos de 2019, 2020 e 2021. Foi denominado Mineração no Brasil.

GRÁFICO 7 – Notícias por Cluster Contexto Geral



Fonte: elaborado pelo autor.

O *Cluster 2* possui 765 notícias, cerca de 23,73% do total, e é caracterizado pela temática Bolsa de Valores, por ter palavras com maior ocorrência como ações, ibovespa e vale. Esse conjunto de palavras aponta para notícias sobre resultados anuais e trimestrais que empresas atuantes da bolsa de valores de São Paulo (IBOVESPA) devem apresentar. Destaca-se a CSN, empresa do ramo siderúrgico que possui plantas de mineração. Em termos de quantidade de notícias por ano, cerca de 50% estão concentradas nos anos de 2019, 2020 e 2021, mostrando ser esse o discurso padrão ao longo dos anos. Foi denominado Ibovespa.

O *Cluster 3* possui cerca de 103 notícias, ou 3,19% do total. Ao analisar as palavras mais representativas desse *Cluster*, as notícias tratam do preço negociado da tonelada em iunes, do contrato, do giro e da entrega. É um *Cluster* isolado no gráfico por tratar de um discurso muito específico. As notícias desse *Cluster* são apenas dos anos de 2020 e 2021, mostrando a importância dessa temática dentro das notícias coletadas. Foi denominado Mercado chinês presente.

O *Cluster 4* possui cerca de 72 notícias, ou 2,23% do total. Ao analisar as palavras mais representativas, as notícias tratam dos preços dos contratos futuros do minério de ferro na bolsa de Dalian na China, que contém 62% de teor de ferro e é considerado fino, ou seja, pelo menos 90% do

carregamento é composto por dez milímetros ou menos. É um *Cluster* isolado no gráfico por tratar de um discurso muito específico. As notícias desse *Cluster* são apenas dos anos de 2020 e 2021, mostrando a importância dessa temática dentro das notícias coletadas e o quanto esse *Cluster* diverge dos demais. Foi denominado Mercado chinês futuro.

O *Cluster 5* possui cerca de 127 notícias, ou 3,94% do total. Ao analisar as palavras mais representativas, as notícias tratam basicamente dos preços dos contratos de minério de ferro e da taxa de câmbio usada na conversão do dólar. Esse também é um *Cluster* isolado no gráfico por tratar de um discurso muito específico. As notícias desse *Cluster* são apenas dos anos de 2020 e 2021, mostrando ser um assunto de maior impacto nesses anos, e, consequentemente, divergindo dos demais *Clusters*. Foi denominado Mercado internacional atual.

O contexto Geral possui 60% das notícias concentradas nos anos de 2019, 2020 e 2021, mostrando que as notícias mais atuais possuem uma influência maior ao construir os *Clusters*. Pelo fato dos *Clusters 1* e *2* terem a maior quantidade de notícias e tratarem de assuntos semelhantes, eles possuem forte sobreposição, influenciando o resultado da pontuação *Silhouette*. Os *Clusters 3, 4* e *5* contam uma história ligada ao comércio exterior de minério de ferro, principalmente com a China. As palavras mais representativas por *Cluster* do contexto Geral estão descritas no Quadro 7, abaixo:

QUADRO 7 – As 10 Principais Palavras do Contexto Geral por *Cluster*

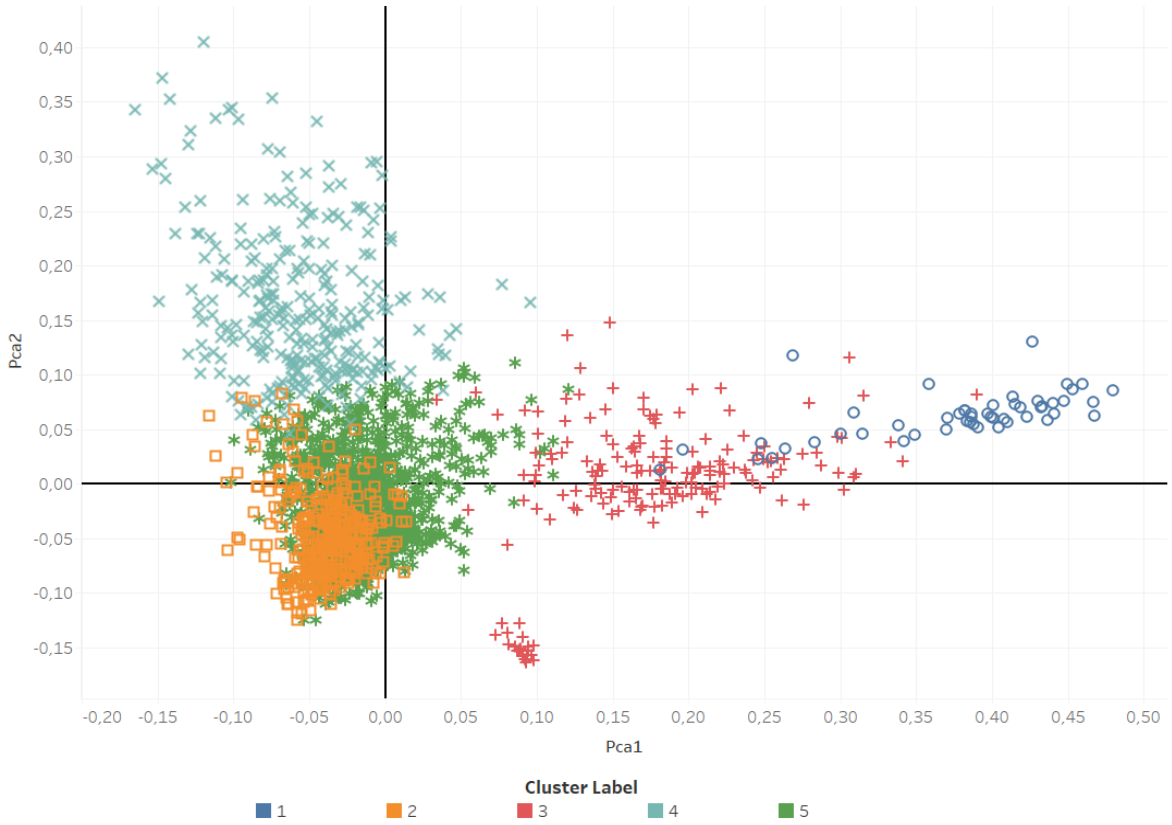
Palavras	Clusters				
	1	2	3	4	5
Palavra 01	mineração	ano	contrato	menos	contratos
Palavra 02	empresa	ações	tonelada	dalian	preços
Palavra 03	brasil	bilhões	iuanes	bolsa	ferro
Palavra 04	vale	minério	giro	ferro	minério
Palavra 05	minas	trimestre	entrega	minério	conversão
Palavra 06	projeto	milhões	negociado	fino	usada
Palavra 07	mineradora	ibovespa	enquanto	milímetros	divulgada
Palavra 08	setor	vale	preço	carregamento	câmbio
Palavra 09	*mining*	alta	dalian	contém	dólar
Palavra 10	mina	csn	futuros	composto	taxa

Fonte: elaborado pelo autor.

No contexto Sem Similaridade, também foram gerados 5 *Clusters* e criados dois eixos por meio da Análise de Componentes, conforme Gráfico 8. A pontuação *Silhouette* foi de 0,0052, evidenciando uma forte sobreposição entre os *Clusters*. Os dados do Gráfico 8 mostram uma sobreposição entre as notícias do *Cluster 2* e *5*, sendo esses *Clusters* os mais representativos, por congregarem cerca de 77% das notícias. As notícias desses *Clusters* estão mais próximas umas das outras, enquanto os demais *Clusters* possuem uma dispersão maior, compartilhando características com os demais. O *Cluster 3* compartilha características com o *Cluster 5* e com o *1* e possui um bloco à parte. Esse tipo de resultado

do gráfico, junto à pontuação *Silhouette,* indica que quando a similaridade semântica é retirada, e o discurso entre os *Clusters* se aproximam ainda mais.

GRÁFICO 8 – Notícias por *Cluster* Contexto Sem Similaridade



Fonte: elaborado pelo autor.

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

O *Cluster 2* possui 352 notícias, cerca de 15,21% do total, e é caracterizado pela temática Agência Nacional de Mineração (ANM), pela barragem e pelo 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, das ações e dos acompanhamentos que foram tomados após o rompimento da barragem e da ameaça que outras barragens representavam. Dado o conjunto de palavras desse *Cluster* e do *Cluster 5,* justifica-se, em parte, o fato de ambos terem forte sobreposição. Foi denominado Barragem.

O *Cluster 3* possui cerca de 180 notícias, ou 7,78% do total. Ao analisar as palavras mais representativas, as notícias tratam basicamente do comportamento da Bolsa de Valores de São Paulo, das variações, tanto de alta quanto de baixa, dos ganhos e das 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 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, nas quantidades de milhões e de bilhões, pela empresa Vale. Também é abordada a produção de ouro, que é o segundo mineral mais extraído no Brasil. A correlação que há entre esse *Cluster*, o *5* e o *2* se dá 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, pois, como se trata de um assunto que deve ser divulgado, ocorre ao longo dos anos. Foi denominado Anúncio da produção.

O *Cluster 5* é caracterizado pela quantidade de notícias, cerca de 1.441 ou 62,25% do total. Ao analisar as palavras mais representativas, as notícias tratam do setor e do mercado de mineração no Brasil e da empresa mineradora Vale. Também é abordada a questão de projeto. Basicamente, a maioria das notícias traz essas palavras, o que justifica esse ser o maior *Cluster*. Em termos de quantidade de notícias por ano, cerca de 52,48% estão concentradas nos anos de 2019, 2020 e 2021, mostrando a influência das notícias dessas datas no *Cluster*, como também o fato de esses serem os anos com maior quantidade de notícias. Além disso, essas são palavras que ocorrem ao longo os demais anos. Foi denominado Mineração no Brasil.

O contexto Sem Similaridade possui 52,48% das notícias concentradas nos anos de 2019, 2020 e 2021, mostrando que as mais atuais possuem uma influência maior ao construir os *Clusters*. Pelo fato dos *Clusters 2* e *5* terem a maior quantidade de notícias e tratarem de assuntos semelhantes, eles possuem forte sobreposição, influenciando o resultado da pontuação *Silhouette*. Os *Clusters 1, 3* e *4*, por estarem próximos, guardam semelhança com os demais e entre si e, mesmo assim, tratam de assuntos diferentes. Os *Clusters* do contexto Sem Similaridade 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 à mineração, que possuem relevância para serem noticiados. As palavras mais representativas por *Cluster* do contexto Geral estão descritas no Quadro 8, abaixo:

QUADRO 8 – As 10 Principais Palavras do Contexto Sem Similaridade por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	tóquio	mineração	ibovespa	minério	mineração
Palavra 02	bolsa	barragem	pontos	ferro	empresa
Palavra 03	registradas	minas	índice	ano	*mining*
Palavra 04	fechou	vale	*brazilian*	trimestre	brasil
Palavra 05	valores	barragens	alta	milhões	projeto
Palavra 06	paulo	nacional	baixa	produção	companhia
Palavra 07	maiores	agência	ganhos	toneladas	setor
Palavra 08	ações	estado	ações	bilhões	mineradora
Palavra 09	altas	anm	queda	ouro	vale
Palavra 10	sessão	região	*machinery*	vale	mercado

Fonte: elaborado pelo autor.

Ao comparar o contexto Geral com o Sem Similaridade, os dados mostram dois *Clusters* que possuem o mesmo discurso, sendo eles os *Clusters* denominados Maior Concentração e Ibovespa. Os demais *Clusters* mudam a narrativa. No contexto Geral, tratam do mercado e da bolsa de valores chineses, e no contexto Sem Similaridade, tratam de fatos internos ao mercado brasileiro, como o rompimento de barragem e a produção, e da bolsa japonesa. Os dados também mostram a influência do tempo nos contextos, especialmente na análise dos *Clusters*, pois alguns são mais influenciados por notícias antigas, enquanto outros, por notícias mais recentes.

4.3 Análise Longitudinal dos Resultados no Contexto Sem Similaridade

Nessa seção, serão analisadas as sumarizações das notícias por ano, apenas no contexto Sem Similaridade, visto que, no contexto Geral, devido à repetição semântica das notícias, há uma redundância nos termos, camuflando, de certa forma, o teor das notícias com maior potencial de recuperação e de uso da informação. A análise por ano se mostra necessária para captar a evolução das narrativas e dos discursos, de forma que um ano ou um conjunto de anos não contamine toda a análise, como observado na seção anterior. As notícias em todos os anos foram agrupadas em 5 *Clusters* e o gráfico desses pode ser visualizado acessando a ferramenta on-line.

A partir de 2018, quando há mais notícias, o indicador *Silhouette* reduz ano a ano, indicando uma sobreposição entre os *Clusters* e um compartilhamento de um mesmo grupo de palavras cada vez maiores, porém compondo discursos que abordam aspectos relacionados ao setor de mineração. A partir de 2019, os *Clusters* passam a captar

subtemáticas sendo necessária uma classificação multi facetada para representar a informação.

4.3.1 Análise do Contexto Sem Similaridade do ano de 2013

No ano de 2013, foram captadas 116 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0112, evidenciando uma forte sobreposição entre os *Clusters*. Os *Clusters 2* e *5* se sobrepõem e parte desses se propõe ao *Cluster 1*, enquanto o *Cluster 3* se encontra misturado entre eles. O *Cluster 4* não apresenta sobreposição com nenhum outro *Cluster*, indicando se tratar de um assunto diferente dos demais. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 9, abaixo:

QUADRO 9 – As 10 Principais Palavras do Contexto Sem Similaridade de 2013 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	empresa	ogx	libertação	mmx	ibovespa
Palavra 02	mineração	queda	eln	minério	alta
Palavra 03	ouro	mercado	tibete	toneladas	valorização
Palavra 04	projeto	ações	exército	ferro	ganhos
Palavra 05	marco	ibovespa	bogotá	recursos	brasília
Palavra 06	ano	presse	solar	azul	horário
Palavra 07	anos	baixa	*wober*	serra	pontos
Palavra 08	belo	*the*	dias	bilhões	ações
Palavra 09	*mining*	índice	serviços	milhões	*shares*
Palavra 10	país	recuo	nacional	sudeste	forte

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 33 notícias, ou 28,45% do total. Ao analisar as palavras mais representativas, as notícias tratam basicamente de empresas de mineração de ouro no país. O termo *mining* aparece devido ao nome das empresas internacionais de ouro que, nesse ano, anunciaram projetos de mineração de ouro. Foi denominado Mineração de ouro.

O *Cluster 2* possui cerca de 37 notícias, ou 31,90% do total. Ao analisar as palavras mais representativas, as notícias tratam do recuo, da baixa, da queda, do mercado, do índice, das ações, da ibovespa e da empresa OGX. Dada a relevância do Cluster devido à quantidade de notícias agrupadas, pode-se deduzir a importância dessa empresa no índice da IBOVESPA. A sobreposição desse Cluster com o Cluster 5 se dá por compartilharem os termos ibovespa e ações. Foi denominado Bolsa e OGX.

O *Cluster 3* possui cerca de 5 notícias, ou 4,31% do total. Essas 5 notícias não possuem correlação com mineração e podem ser consideradas como erro na pesquisa. Foi denominado Sem correlação.

O *Cluster 4* possui cerca de 15 notícias, ou 12,93% do total. Por estar mais afastado dos demais e por não haver sobreposição, pode-se deduzir que trata de outro assunto. Isso é observável ao analisar as palavras mais representativas, pois as notícias tratam da produção de bilhões e milhões de toneladas minério de ferro na unidade de Serra Azul pela empresa MMX. O termo chave desse *Cluster* é a palavra MMX, pois serve de ligação entre as notícias e caracteriza um *Cluster* que não guarda correlação com os demais. Foi denominado MMX.

O *Cluster 5* possui cerca de 26 notícias, ou 22,41% do total. Possui sobreposição com o *Cluster 2* por tratar de variação da IBOVESPA. A diferença entres esses está no fato do *Cluster 5* tratar da valorização e dos ganhos da IBOVESPA enquanto o *Cluster 2* trata de perdas ligadas a empresa OGX. Foi denominado Ibovespa.

Os *Clusters* do contexto Sem Similaridade do ano de 2013 tratam basicamente de projetos de mineração de minério de ferro e de ouro, da queda das ações da empresa OGX e de seu efeito na IBOVESPA, da variação positiva da IBOVESPA quando a OGX não está associada e da produção de toneladas de minério de ferro pela MMX na unidade de Serra Azul.

4.3.2. Análise do Contexto Sem Similaridade do ano de 2014

No ano de 2014, foram captadas 183 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,009, evidenciando uma forte sobreposição entre os *Clusters*. O *Cluster 2 e 5* se sobrepõem e parte desses se propõe ao *Cluster 1* enquanto o *Cluster 3* se encontra misturado entre eles. O *Cluster 4* não apresenta sobreposição com nenhum outro *Cluster*, indicando que trata de um assunto diferente dos demais. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 10, abaixo:

QUADRO 10 – As 10 Principais Palavras do Contexto Sem Similaridade de 2014 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	ibovespa	ano	ibovespa	minério	projetos
Palavra 02	alta	empresa	*fell*	ferro	empresa
Palavra 03	desvalorização	mineração	*list*	preço	mil
Palavra 04	mmx	ouro	ibovespas	disse	brasil
Palavra 05	valorização	milhões	*shares*	produção	mineração
Palavra 06	baixas	projeto	*losses*	companhia	ouro
Palavra 07	recuo	*mining*	*depreciation*	mercado	hoje
Palavra 08	maiores	segundo	baixa	indústria	projeto
Palavra 09	altas	mercado	*decline*	depósito	milhões
Palavra 10	petrobras	mina	*drop*	preços	acordo

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 33 notícias, ou 18,03% do total. Ao analisar as palavras mais representativas, as notícias tratam da desvalorização das ações da MMX e da Petrobras na IBOVESPA. A sobreposição com o *Cluster 3* ocorre devido ao compartilhamento do termo IBOVESPA. Foi denominado MMX.

O *Cluster 2* possui cerca de 63 notícias, ou 34,43% do total. Ao analisar as palavras mais representativas, as notícias tratam dos projetos das empresas de mineração de ouro que possuem o termo *mining* em seus nomes. Esse *Cluster* guarda correlação com o *Cluster 5* por tratar de projetos de mineração de ouro. Foi denominado Mineração de ouro – empresas.

O *Cluster 3* possui cerca de 27 notícias, ou 14,75% do total. Essas notícias tratam das perdas da IBOVESPA, mas no idioma inglês. Foi denominado Ibovespa.

O *Cluster 4* possui cerca de 23 notícias, ou 12,57% do total. Por estar mais afastado dos demais e por não haver sobreposição, pode-se deduzir que tratar de outro assunto. Isso é observado ao analisar as palavras mais representativas, pois tratam do preço do minério de ferro e de depósitos de minério de ferro. Foi denominado Preço do minério de ferro e estoque de minério de ferro.

O *Cluster 5* possui cerca de 37 notícias, ou 20,22% do total. Possui sobreposição com o *Cluster 2* por tratar de projetos de mineração de ouro no Brasil. Foi denominado Mineração de ouro.

Os *Clusters* do contexto Sem Similaridade do ano de 2014 tratam basicamente de projetos de mineração de minério de ouro, das oscilações do índice da IBOVESPA, seja por causa da MMX ou da Petrobras, e da questão do preço do minério de ferro.

4.3.3. Análise do Contexto Sem Similaridade do ano de 2015

No ano de 2015, foram captadas 167 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0068, evidenciando uma forte sobreposição entre os *Clusters*. O *Cluster 3* se encontra diluído no *Cluster 2* e esse compartilha a borda de decisão com o *Cluster 4* e *5* enquanto o *Cluster 4* compartilha a borda de decisão com os *Clusters 1* e *5*. Esse tipo de resultado mostra os *Clusters* tratando dos mesmos assuntos, mas com alguma variação. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 11, abaixo:

QUADRO 11 – As 10 Principais Palavras do Contexto Sem Similaridade de 2015 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	jaguar	empresa	brasil	novo	minério
Palavra 02	trimestre	capital	mineração	empresa	ferro
Palavra 03	ouro	mineração	setor	ceo	toneladas
Palavra 04	ano	produtividade	país	companhia	milhões
Palavra 05	onças	gestão	mineral	conselho	preço
Palavra 06	sonhos	segurança	inovação	executivo	queda
Palavra 07	castelo	global	investimentos	presidente	preços
Palavra 08	turmalina	pesquisa	exploração	*hexagon*	ano
Palavra 09	terceiro	forma	desenvolvi-mento	mineração	mercado
Palavra 10	emissão	relatório	evento	diretor	ouro

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 15 notícias, ou 8,98% do total. Ao analisar as palavras mais representativas, as notícias tratam da produção de ouro, no ano e no trimestre, da empresa *Jaguar Mining*, mineradora de ouro. Os termos castelo e sonhos são referentes a um projeto de mineração de ouro com esse nome, Castelo dos Sonhos. Foi denominado Mineração de ouro – empresas.

O *Cluster 2* possui cerca de 34 notícias, ou 20,36% do total. Ao analisar as palavras mais representativas, as notícias tratam de diversos assuntos ligados à produtividade, à gestão, à pesquisa e à segurança da mineração. A sobreposição com o *Cluster 3* é proveniente dessa temática. Foi denominado Pesquisa e Desenvolvimento.

O *Cluster 3* possui 34 notícias, ou 20,6% do total, e trata de investimentos em inovação para exploração do setor mineral e de mineração no Brasil/país. Foi denominado Inovação.

O *Cluster 4* possui cerca de 42 notícias, ou 25,15% do total. Ele trata da troca de dirigentes, seja no conselho, seja na direção das empresas de mineração. Foi denominado Mudança de diretoria.

O *Cluster 5* também possui 42 e aborda a produção de milhões de toneladas de minério de ferro e de ouro e a queda de seus preços. Foi denominado Anúncio de produção, Preço do minério de ferro e Preço do ouro.

Os *Clusters* desse contexto e ano tratam basicamente da produção de ouro, de temas ligados à inovação e à pesquisa em mineração no Brasil, da troca de diretores e de conselheiros em empresas de mineração e da oscilação do preço do minério de ferro e do ouro.

4.3.4 Análise do Contexto Sem Similaridade do ano de 2016

No ano de 2016, foram captadas 235 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0056, evidenciando uma forte sobreposição entre os *Clusters*. Apesar disso, as notícias estão dispersas, fazendo os *Clusters* compartilharem as bordas. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 12, abaixo:

QUADRO 12 – As 10 Principais Palavras do Contexto Sem Similaridade de 2016 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	mineração	queda	novo	ouro	minério
Palavra 02	setor	pontos	conselho	projeto	ferro
Palavra 03	mineral	baixa	empresa	milhões	preço
Palavra 04	*mining*	após	companhia	sondagem	porto
Palavra 05	empresa	índice	brasil	mina	produção
Palavra 06	*world*	tóquio	cargo	mineradora	toneladas
Palavra 07	*congress*	alta	executivo	onças	pedreira
Palavra 08	indústria	fechou	presidente	mil	china
Palavra 09	janeiro	vez	administração	ano	vale
Palavra 10	brasil	bolsa	diretor	campanha	tinto

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 71 notícias, ou 30,21% do total. Ao analisar as palavras mais representativas, as notícias tratam do congresso WMC (*World Mining Congress*), que ocorreu no estado do Rio de Janeiro. Foi denominado Congresso WMC.

O *Cluster 2* possui cerca de 39 notícias, ou 16,60% do total e aborda a oscilação do índice da bolsa de Tóquio. Foi denominado Bolsa de Tóquio.

O *Cluster 3* possui cerca de 45 notícias, ou 19,15% do total. Trata da mudança de dirigentes nas empresas e no conselho. Por ser um fato relevante, as empresas devem comunicá-lo ao mercado para manter os investidores informados. Foi denominado Mudança de diretoria.

O *Cluster 4* possui cerca de 53 notícias, ou 22,55% do total, e aborda a sondagem de minas para projeto de mineração de ouro e as quantidades em onças. Foi denominado Mineração de ouro.

O *Cluster 5* possui cerca de 27 notícias, ou 11,49% do total. Trata da produção e do preço do minério de ferro. Foi denominado Anúncio de Produção e Preço do minério de ferro.

Os *Clusters* desse contexto e ano compartilham um mesmo grupo de palavras e a ordem dessas palavras diferencia os *Clusters*. Mesmo assim, o discurso segue abordando características mais relevantes da mineração nesse ano, com destaque para notícias de mineração de ouro e para a troca de comando das empresas.

4.3.5 Análise do Contexto Sem Similaridade do ano de 2017

No ano de 2017, foram captadas 217 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0366, evidenciando uma forte sobreposição entre os *Clusters*. Os *Clusters 1, 3* e *5* se sobrepõem, indicando tratarem do mesmo tema, enquanto os *Clusters* 2 e 4 não apresentam sobreposição com nenhum outro *Cluster*, indicando que tratam de um assunto diferente dos demais. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 13, abaixo:

QUADRO 13 – As 10 Principais Palavras do Contexto Sem Similaridade de 2017 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	ouro	*brazilian*	tóquio	eike	mineração
Palavra 02	ano	*machinery*	bolsa	cabral	empresa
Palavra 03	produção	*brazil*	sessão	batista	setor
Palavra 04	mineração	*program*	paulo	rio	projetos
Palavra 05	brasil	*abimaq*	pontos	sérgio	além
Palavra 06	mina	*promotion*	ltd	empresário	*mining*
Palavra 07	empresa	*solutions*	alta	milhões	votorantim
Palavra 08	projeto	*equipment*	índice	federal	metais
Palavra 09	toneladas	*agency*	fechou	operação	austrália
Palavra 10	mineradora	*exports*	ganhos	janeiro	programa

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 120 notícias, ou 55,30% do total, e concentra a maior parte delas, o que justifica, em parte, os *Clusters 3* e *5* possuírem semelhanças com ele. As palavras mais representativas desse *Cluster* tratam de notícias que abordam projetos de mineração e produção de toneladas de ouro. Foi denominado Anúncio de produção e Mineração de ouro.

O *Cluster 2* possui cerca de 17 notícias, ou 7,83% do total. Ele aborda notícias no idioma inglês que tratam da promoção de soluções de equipamentos e de maquinário para mineração, proporcionados pela ABIMAQ – Associação Brasileira de Máquinas. Foi denominado ABIMAQ.

O *Cluster 3* possui cerca de 13 notícias, ou 5,99% do total, e trata da oscilação do índice da bolsa de Tóquio. A sobreposição com o *Cluster 1* ocorre pelo compartilhamento de palavras-chaves, como mineração e empresa. Foi denominado Bolsa de Tóquio.

O *Cluster 4* possui cerca de 19 notícias, ou 8,76% do total, e aborda uma operação da Polícia Federal que envolveu o empresário Eike Batista, dono da MMX e OGX, e o governador do Rio de Janeiro na época, Sérgio Cabral. Essa notícia foi captada por envolver a MMX, empresa de mineração. Foi denominado MMX.

O *Cluster 5* possui cerca de 48 notícias, ou 22,12% do total, e trata da relação de empresas siderúrgicas com empresas de mineração. Foi denominado Empresas siderúrgicas/mineradoras.

Os *Clusters* desse contexto e desse ano compartilham a mesma temática, diferenciando-se em poucos aspectos, especialmente os *Clusters 1, 3* e *5*. Nesse contexto, também há um destaque para notícias de projetos de produção de ouro.

4.2.6 Análise do Contexto Sem Similaridade do ano de 2018

No ano de 2018, foram captadas 182 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0178, evidenciando uma forte sobreposição entre os *Clusters*. Apesar dos *Clusters* guardarem correlação, exceto o *Cluster 1*, que trata de um assunto diferente, os demais abordam variações de temática de mineração. Nesse ano, não há um *Cluster* com mais de 50% das notícias, logo, um *Cluster* não atua como influenciador sobre os demais, como ocorre nos *Clusters* do ano de 2017, por exemplo. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 14, abaixo.

QUADRO 14 – As 10 Principais Palavras do Contexto Sem Similaridade de 2018 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	registradas	empresa	milhões	mineração	produção
Palavra 02	tóquio	bilhões	cfem	nexa	mil
Palavra 03	valores	mundo	minerais	projetos	trimestre
Palavra 04	bolsa	maior	recursos	gestão	ano
Palavra 05	ações	mineração	mineração	*startups*	toneladas
Palavra 06	maiores	*mining*	estado	áreas	onças
Palavra 07	paulo	brasil	exploração	brasil	minério
Palavra 08	altas	disse	rio	setor	ouro
Palavra 09	sessão	cobre	compensação	evento	mina
Palavra 10	fechou	valor	*royalties*	tecnologia	aumento

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 11 notícias, ou 6,04% do total, e aborda as oscilações do índice da bolsa de valores de Tóquio. Foi denominado Bolsa de Tóquio.

O *Cluster 2* possui cerca de 65 notícias, ou 35,71% do total, sendo o maior *Cluster* desse ano em quantidade de notícias. Ele aborda as maiores empresas de mineração no Brasil e no mundo e abrange a mineração de cobre. Foi denominado Empresas de mineração e Mineração de cobre.

O *Cluster 3* possui cerca de 17 notícias, ou 9,34% do total, e aborda a CFEM (Compensação Financeira pela Exploração Mineral), que é um imposto pago pelas empresas mineradoras em forma de *Royalties* para os Municípios e para os órgãos da administração da União, como contraprestação pela utilização econômica dos recursos minerais em seus respectivos territórios. Foi denominado CFEM.

O *Cluster 4* possui cerca de 48 notícias, ou 26,37% do total, e aborda a promoção de eventos organizados pela empresa *Nexa Resources* para investir em *startups* que desenvolvem projetos de tecnologia para o setor de mineração. Foi denominado Inovação.

O *Cluster 5* possui cerca de 41 notícias, ou 22,53% do total. Ele aborda o aumento da produção de ouro e de minério de ferro no trimestre e no ano. Foi denominado Anúncio de produção.

Os *Clusters* desse contexto e ano compartilham a mesma temática, diferenciando-se em aspectos peculiares. Nesse ano, novas temáticas importantes para o setor de mineração foram abordadas, como a CFEM, investimento em *startups* voltadas para mineração e para mineração de cobre.

4.3.7 Análise do Contexto Sem Similaridade do ano de 2019

No ano de 2019, foram captadas 296 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0082, evidenciando uma forte sobreposição entre os *Clusters*. Os *Clusters* guardam correlação, exceto o *Cluster 4*, que trata de um assunto diferente dos demais, e o *Cluster 1*, que compartilha a borda com os demais. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 15, abaixo:

QUADRO 15 – As 10 Principais Palavras do Contexto Sem Similaridade de 2019 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	vale	mineração	milhões	fechou	soluções
Palavra 02	barragem	energia	toneladas	valores	nova
Palavra 03	mineradora	projeto	ano	tóquio	*mining*
Palavra 04	brumadinho	empresa	produção	registradas	empresa
Palavra 05	rompimento	governo	trimestre	bolsa	startups
Palavra 06	barragens	minas	período	altas	mineração
Palavra 07	empresa	presidente	valor	maiores	anos
Palavra 08	após	indígenas	pará	batista	*green*
Palavra 09	rejeitos	setor	minas	eike	brasil
Palavra 10	mineração	meio	vale	ações	startup

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 76 notícias, ou 25,68% do total, e aborda o rompimento da barragem de rejeitos da Vale, em Brumadinho. Foi denominado Barragem.

O *Cluster 2* possui cerca de 112 notícias, ou 37,84% do total, sendo o maior Cluster desse ano em quantidade de notícias. Pelas palavras mais representativas, é possível analisar subtemáticas, sendo uma ligada a projetos de energia para mineração, que envolve o governo do estado de Minas Gerais, e a outra subtemática é a demarcação e a mineração em terras indígenas. Foi denominado Energia e Mineração em terras indígenas.

O *Cluster 3* possui cerca de 23 notícias, ou 7,77% do total, e aborda a produção de toneladas de minério de ferro pela empresa Vale em suas unidades no Pará e em Minas Gerais. No entanto, ao recuperar os resumos, a empresa Vale anunciou o aumento da produção no estado do Pará, e não nas unidades de Minas Gerais, indicando uma mudança estratégica frente à tragédia em Brumadinho. Foi denominado Anúncio de Produção e Mudança de estratégia.

O *Cluster 4* possui cerca de 17 notícias, ou 5,74% do total, e aborda a oscilação do índice da bolsa de Tóquio, além de citar o empresário Eike Batista em uma outra subtemática. Foi denominado Bolsa de Tóquio e MMX.

O *Cluster 5* possui cerca de 68 notícias, ou 22,97% do total. Ele aborda ações promovidas pela *Startup Green Mining*, que desenvolve soluções para o setor de mineração. Foi denominado Inovação.

Os *Clusters* desse contexto e desse ano compartilham a mesma temática e se diferenciam em subtemáticas, que, por sua vez, abordam novos assuntos relevantes para o setor, como o aspecto de legislação quando o governo federal propõe possibilitar a mineração em terras indígenas.

4.3.8 Análise do Contexto Sem Similaridade do ano de 2020

No ano de 2020, foram captadas 413 notícias sem similaridade semântica. A pontuação *Silhouette* foi de 0,0061, evidenciando uma forte sobreposição entre os *Clusters*. Apesar dos *Clusters* guardarem correlação, exceto o *Cluster 5*, que trata de um assunto diferente, os demais *Clusters* compartilham características semelhantes, principalmente o *Cluster 4*. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadro 16, abaixo:

QUADRO 16 – As 10 Principais Palavras do Contexto Sem Similaridade de 2020 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	reum	mineração	mineração	vale	tóquio
Palavra 02	rede	energia	vale	milhões	fechou
Palavra 03	mineradores	indígenas	minério	trimestre	registradas
Palavra 04	taxas	governo	ferro	bilhões	valores
Palavra 05	pessoas	nacional	empresa	ano	bolsa
Palavra 06	atualização	projeto	barragens	empresa	maiores
Palavra 07	blocos	desenvolvimento	barragem	mineradora	paulo
Palavra 08	transação	programa	setor	mineração	pregão
Palavra 09	gás	presidente	mineradora	paulo	ações
Palavra 10	actros	setor	rio	ouro	altas

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 21 notícias, ou 5,08% do total, e trata de mineração de *bitcoin* e de um caminhão desenvolvido para operar em mineradoras. Em relação às reportagens sobre *bitcoin*, é importante ressaltar a capacidade dos algoritmos de separar os

assuntos, evitando misturar aqueles que não são foco da pesquisa. Foi denominado Sem correlação e Inovação.

O *Cluster 2* possui cerca de 81 notícias, ou 19,61% do total. Pelas palavras mais representativas, é possível identificar subtemáticas, sendo: a) o desenvolvimento de um programa para o setor de mineração, elaborado pelo governo federal; b) notícias ligadas ao fornecimento de energia para o setor de mineração; c) mineração em terras indígenas. Foi denominado Apoio governamental, Energia e Mineração em terras indígenas.

O *Cluster 3* possui cerca de 174 notícias, ou 42,13% do total, sendo o maior *Cluster* deste ano em quantidade de notícias. Esse *Cluster* aborda a questão das barragens de minério de ferro. Como o *Cluster* possui palavras-chaves, como mineração, vale, minério e empresa, ele compartilha semelhanças com os demais *Clusters*, justificando a sobreposição entre eles. Foi denominado Barragem.

O *Cluster 4* possui cerca de 125 notícias, ou 30,27% do total, e aborda a produção de minério pela empresa Vale, no ano e no trimestre, em quantidade de milhões e bilhões, e cita a mineração de ouro, em questão de período e de quantidade. Foi denominado Anúncio de produção e Mineração de ouro.

O *Cluster 5* possui cerca de 12 notícias, ou 2,91% do total, e aborda oscilações na bolsa de valores de Tóquio. Foi denominado Bolsa de Tóquio.

Os *Clusters* desse contexto e desse ano compartilham a mesma temática e se diferenciam em subtemáticas. Os dois maiores *Clusters*, *3* e *4*, não trazem questões relevantes e novas para o debate sobre mineração.

4.3.9 Análise do Contexto Sem Similaridade do ano de 2021

No ano de 2021, foram captadas 506 notícias sem similaridade semântica, distribuídas em 5 *Clusters*. A pontuação *Silhouette* foi de 0,005, evidenciando uma forte sobreposição entre os *Clusters*. Apesar de os *Clusters* guardarem correlação, exceto o *Cluster 5,* que trata de um assunto diferente, os demais compartilham características semelhantes. As palavras mais representativas por *Cluster* nesse contexto e nesse ano estão descritas no Quadros 17 abaixo:

QUADRO 17 – As 10 Principais Palavras do Contexto Sem Similaridade de 2021 por *Cluster*

Palavras	Clusters				
	1	2	3	4	5
Palavra 01	mineração	trimestre	vale	mineração	índice
Palavra 02	mmx	minério	bilhões	energia	pontos
Palavra 03	administração	csn	mineradora	empresa	ações
Palavra 04	conselho	milhões	empresa	setor	bolsa
Palavra 05	operações	ferro	ações	brasil	alta
Palavra 06	publicado	ano	acordo	áreas	tóquio
Palavra 07	companhia	preços	oferta	projeto	fechou
Palavra 08	estado	mineração	ano	meio	queda
Palavra 09	vale	toneladas	mineração	anos	europeias
Palavra 10	ouro	tonelada	companhia	minas	valores

Fonte: elaborado pelo autor.

O *Cluster 1* possui cerca de 75 notícias, ou 14,82% do total, e trata de três subtemáticas: a) da empresa MMX; b) da troca de membros da administração e conselho de mineradora; c) da empresa Vale; d) da mineração de ouro no estado. Foi denominado MMX e Mudança de diretoria e Mineração de ouro.

O *Cluster 2* possui cerca de 96 notícias, ou 18,97% do total, e traz um novo *player* de mineração, que já havia aparecido nos anos anteriores, mas que, nesse ano, ganha volume em termos de citação. Trata-se da empresa CSN Mineração, do setor siderúrgico, que abriu uma planta de mineração. As notícias abordam a produção de minério de ferro, no ano e no trimestre, em milhões de toneladas. As notícias também citam uma questão de preço. Foi denominado Novo *Player*, Anúncio de produção e Preço do minério de ferro.

O *Cluster 3* possui cerca de 110 notícias, ou 21,74% do total, sendo o segundo mais volumoso, e trata da oferta de ações da mineradora Vale. Pelas palavras mais representativas, é possível observar que há várias formas de se referir à Vale, sendo elas empresa, mineradora e companhia. Foi denominado Oferta de ações.

O *Cluster 4* possui cerca de 194 notícias, ou 38,34% do total, e aborda a questão de projetos de energia para e pelas empresas de mineração. Como algumas palavras usadas nesse *Cluster* são iguais às usadas no *Cluster 3*, é possível compreender o motivo desses *Clusters* estarem sobrepostos. Foi denominado Energia.

O *Cluster 5* possui cerca de 31 notícias, ou 6,13% do total, e aborda oscilações na bolsa de valores de Tóquio. Foi denominado Bolsa de Tóquio.

Os *Clusters* desse contexto e desse ano compartilham a mesma temática e se diferenciam em subtemáticas.

4.4 Mensurar a Repetição de Narrativas ao Longo do Tempo

Dentre os objetivos específicos deste trabalho foi estipulada a mensuração da repetição de narrativas/discursos ao longo do tempo. O foco dessa ação foi compreender se há um tema que se repete em todos os anos, em alguns, ou se há temas isolados, mostrando a importância deles no ano correspondente.

Cada *Cluster* por ano do Contexto Sem Similaridade recebeu um nome, que será Narrativa/Discurso. Cada Narrativa/Discurso foi inserida em uma tabela e foi contado em quantos anos essa Narrativa/Discurso apareceu, conforme a tabela abaixo:

Tabela 2 – Quantidade de Ocorrências de Narrativas/Discurso por ano

Continua

Narrativas/Discurso	Anos									Total Geral
	2013	2014	2015	2016	2017	2018	2019	2020	2021	
Anúncio de produção	0	0	1	1	1	1	1	1	1	7
Bolsa de Tóquio	0	0	0	1	1	1	1	1	1	6
Mineração de ouro	1	1	0	1	1	0	0	1	1	6
MMX	1	1	0	0	1	0	1	0	1	5
Inovação	0	0	1	0	0	1	1	1	0	4
Preço do minério de ferro	0	1	1	1	0	0	0	0	1	4
Energia	0	0	0	0	0	0	1	1	1	3
Mudança de diretoria	0	0	1	1	0	0	0	0	1	3
Barragem	0	0	0	0	0	0	1	1	0	2
Ibovespa	1	1	0	0	0	0	0	0	0	2
Mineração de ouro – empresas	0	1	1	0	0	0	0	0	0	2
Mineração em terras indígenas	0	0	0	0	0	0	1	1	0	2
Sem correlação	1	0	0	0	0	0	0	1	0	2
ABIMAQ	0	0	0	0	1	0	0	0	0	1
Apoio governamental	0	0	0	0	0	0	0	1	0	1
Bolsa e OGX	1	0	0	0	0	0	0	0	0	1
CFEM	0	0	0	0	0	1	0	0	0	1
Congresso WMC	0	0	0	1	0	0	0	0	0	1
Empresas de mineração	0	0	0	0	0	1	0	0	0	1
Empresas siderúrgicas/mineradoras	0	0	0	0	1	0	0	0	0	1
Estoque de minério de ferro	0	1	0	0	0	0	0	0	0	1
Mineração de cobre	0	0	0	0	0	1	0	0	0	1
Mudança de estratégia	0	0	0	0	0	0	1	0	0	1
Novo Player	0	0	0	0	0	0	0	0	1	1
Oferta de ações	0	0	0	0	0	0	0	0	1	1

Tabela 2 – Quantidade de Ocorrências de Narrativas/Discurso por ano

Conclusão

Narrativas/Discurso	Anos									Total Geral
	2013	2014	2015	2016	2017	2018	2019	2020	2021	
Pesquisa e Desenvolvimento	0	0	1	0	0	0	0	0	0	1
Preço do ouro	0	0	1	0	0	0	0	0	0	1

Fonte: elaborado pelo autor.

Esse grupo de notícias captadas tem um enfoque em temas relacionados a questões de bolsas de valores. Isso justifica, em parte, a ocorrência da Narrativa/Discurso “Anúncio de produção” em 7 dos 9 anos. As empresas de mineração listadas em bolsas de valores, como a empresa Vale, devem informar trimestralmente a quantidade de minério de ferro produzido e qual a expectativa de produção até o final do ano. Também foram captadas as variações da Bolsa de Valores de Tóquio quando essa tratava de variações das ações de empresas mineradoras. Entretanto, essa narrativa/discurso não aponta mudanças no macroambiente ou possui capacidade ínfima de o fazer.

A narrativa/discurso “Mineração de ouro” apareceu com destaque em vários anos, pois o ouro é o segundo principal mineral extraído no Brasil. Essa narrativa/discurso envolve basicamente notícias relacionadas a projetos. Essa narrativa/discurso demanda um detalhamento maior dessas notícias para captar nuances capazes de apontar mudanças em algum aspecto do macroambiente. A “MMX” apareceu em 5 anos dos 9 da base, e a saída dessa empresa do mercado acarretou mudanças no aspecto social e ecológica/ambiental, pois as comunidades nas quais a empresa possuía plantas ou projetos foram afetadas. A “Inovação” se mostrou um tema de maior importância entre os anos 2018 e 2020, já que há mais notícias ligadas a ele. Parte da construção dessa narrativa/discurso é provocada pelos avanços de *Machine Learning* e de *Deep Learning*, além de pesquisa de novos materiais. O “preço do minério de ferro” ocorreu nos anos de 2014 a 2016 e voltou em 2021, mostrando ser um tema cíclico e, dada a repetição, surgiu como uma narrativa/discurso a ser monitorado.

As narrativas/discursos “Energia” e “Mudança de diretoria” ocorreram apenas em 3 anos dos 9, mas com uma distinção muito clara. Energia sempre foi um tema prioritário para mineração e esse tema apareceu apenas em 2019, 2020 e 2021, indicando uma mudança na forma de discutir energia em mineração. Mudança de diretoria é algo que ocorre com frequência, mas, nesses anos, foram eventos com maior propagação, resultando em um tema de destaque.

As narrativas/discursos com apenas duas ocorrências apontam temas que surgiram em um ano e tiveram forte reverberação no ano seguinte enquanto as que tiveram apenas uma ocorrência apontam para temas isolados, de forte destaque no ano. O aparecimento desses temas indica ser necessário algum tipo de acompanhamento, exceto em relação àqueles muito pontuais, como “Bolsa e OGX” e “Congresso WMC”.

5 RESULTADOS DA AVALIAÇÃO DO MODELO CONSTRUÍDO

Ao todo, foram entrevistados 15 especialistas, sendo 4 de Mineração e 11 de Dados e Informação, entre os dias 24/11/2023 e 07/12/2023. Dos 15 entrevistados, apenas 2 tinham especialização em ambas as áreas, porém foram enquadrados na área que possuíam maior tempo de atuação. Cada especialista foi entrevistado após analisar o *Dashboard*, que consta no seguinte *link*: https://luanderfalcao.wixsite.com/tese-de-doutorado.

O roteiro de perguntas foi elaborado para captar as características básicas do entrevistado, consistindo em questões de 1 à 5. As questões de 6 e 7 compreendem o que se busca e onde se busca informação sobre setores, e as questões de 8 a 11 captam a validação da metodologia.

Em termos de caracterização dos entrevistados, eles possuem, em média, 11 anos em suas respectivas áreas, seja de Mineração ou de Dados e Informação, com mediana de 10 anos, indicando que há baixa variabilidade entre o tempo de experiência dos entrevistados.

Cerca de 60% dos entrevistados possuem entre 36 e 45 e, desse total, 89% possuem cargos de especialista até diretor. Do total de entrevistados, cerca de 80% possuem cargos de especialista até diretor, conforme Tabela 3. Esse resultado evidencia a experiência de vida e profissional, principalmente quando se consideram os cargos dos entrevistados.

Tabela 3 – Entrevistados por Faixa Etária e por Cargo Atual

Cargo Atual	Faixa Etária			Total Geral
	Entre 25 e 35	Entre 36 e 45	Acima de 45	
Diretor	0	2	1	3
Gerente	0	1	0	1
Coordenador	1	2	0	3
Especialista	1	3	1	5
Sênior	0	1	0	1
Pleno	2	0	0	2
Total Geral	4	9	2	15

Fonte: elaborado pelo autor.

Quando perguntados sobre quais fontes utilizavam para acompanhar ou se informar quanto às questões setoriais, cerca de 22,22% citaram outras fontes. Cerca de 33% se informam por meio de Relatórios/informativos internos e *Sites* de setores. Dentre as fontes menos utilizadas estão jornais e *newsletters*, conforme a Tabela 4.

Tabela 4 – Fontes de Informações Citadas

Fontes de Informação	Frequência de Citação	Part. %
Outras fontes	10	22,22%
Relatórios/informativos internos	8	17,78%
Sites de setores	7	15,56%
Colegas ou amigos	6	13,33%
Relatórios de consultorias	6	13,33%
Jornais (exemplo: Folha de São Paulo, Estadão, O Globo, Valor Econômico e semelhantes)	4	8,89%
Newsletters	4	8,89%
Total Geral	45	100,00%

Fonte: elaborado pelo autor.

É importante ressaltar que os entrevistados citaram como outras fontes de informações setoriais as redes sociais, principalmente o X (antigo *Twitter*) e o *LinkedIn*. Em ambos os casos, os entrevistados relataram como ponto central o acompanhamento das pessoas certas, pois elas são fontes de notícias sobre o que está acontecendo no setor de interesse. Um dos entrevistados citou um monitoramento sistemático via ferramentas de notícias do *Google*. Outro entrevistado citou que consome informação pré-selecionada, ou seja, notícias que passaram por uma espécie de curadoria.

Quando perguntados sobre quais temas possuem o costume de acompanhar ou monitorar relacionados a setores produtivos, cerca de 48% responderam Novas tecnologias e Desempenho da economia, e 25% responderam Outros e Parcerias, fusões e aquisições. Esse resultado evidencia uma preocupação maior com a temática economia e tecnologia, por serem esses dois os temas com maior potencial de impacto no curto prazo na maioria dos setores.

Tabela 5 – Temas de Monitoramento Citados

Temas de Monitoramento	Frequência de Citação	Part. %
Novas tecnologias	12	25,00%
Desempenho da economia	11	22,92%
Outros	6	12,50%
Parcerias, fusões e aquisições	6	12,50%
Mudança de estratégia	5	10,42%
Questões ligadas ao meio ambiente e sociais	4	8,33%
Aspectos culturais	2	4,17%
Mudanças na legislação ou normas regulatórias	2	4,17%
Total Geral	48	100,00%

Fonte: elaborado pelo autor.

Os entrevistados que citaram “Outros” detalharam que monitoram temas como inovação, em seu sentido mais amplo, a parte política da legislação, tendências e o posicionamento de mercado. Os temas Ambiente de Negócios e Volume de Produção foram citados, pois proporcionam uma visão mais holística dos setores, a fim de monitorar oportunidades e quais *outliers* estão surgindo. Alguns entrevistados ainda detalharam o tema Novas Tecnologias, já que monitoram a evolução das IAs, o lançamento de novos produtos de tecnologia *SaaS*, os produtos *No Code*, as novas ferramentas de tecnologia e a parte de tecnologia do setor de Manufatura e Automobilístico.

Quando questionados sobre o uso do *Dashboard*, que tangibiliza a metodologia, e sobre o quanto a metodologia era válida para responder à pergunta central, cerca de 73% dos entrevistados acharam que o método de tratamento e de apresentação da informação foi muito útil para ajudar no entendimento do setor. Cerca de 47% entenderam que o método utilizado no *Dashboard* possui muita estrutura para mudar o seu entendimento ou a sua percepção sobre o setor ou de alguma parte dele. Além disso, cerca de 33% entenderam que as 10 principais palavras por *Cluster* por ano e por contexto retrataram bastante os principais fatos ocorridos na mineração naquele ano, segundo a sua percepção.

Tabela 6 – Uso do *Dashboard*

Perguntas do Roteiro	Muito	Médio	Pouco	Não retratou	Total Geral
8 – A partir do uso do *Dashboard*, o quanto você acha que método de tratamento e apresentação da informação foi útil para ajudar no entendimento do setor?	11	4	0	0	15
9 – O método utilizado no *Dashboard* possui estrutura capaz de mudar o seu entendimento ou a sua percepção sobre o setor, ou de alguma parte do setor?	7	7	1	0	15
11 – As 10 principais palavras por *Cluster* por ano e por contexto retrataram os principais fatos ocorridos na mineração naquele ano, segundo a sua percepção?	5	7	2	1	15

Fonte: elaborado pelo autor.

Ao longo da entrevista, à medida que o entrevistado ia narrando a sua experiência no uso do *Dashboard* e apontando as suas percepções, as respostas dessas perguntas eram anotadas. De acordo com os entrevistados, alguns fatores validam as respostas, tais como:

- A quantidade de texto coletado ajuda na validação do método;
- O método utilizado vai do abrangente para o específico, do mapa *mundi* para o *Wase*, por exemplo;
- O *Dashboard* ajuda a compreender sobre o setor e o que ocorreu nele ao longo do tempo;
- O *Dashboard* mostra alinhamento com o que foi publicado na mídia e foi acompanhado, e os *Clusters* estão de acordo com o ocorrido no setor de mineração, refletindo o que ocorreu;
- As narrativas/discursos deixam claro do que se trata o *Cluster*;
- Algumas narrativas, como, por exemplo, a CFEM, são mais importantes para um determinado grupo de *Stakeholders*. Nesse caso, a CFEM é mais importante para o governo em suas três esferas de atuação;
- O *Dashboard* ajudou a entender melhor o setor, mas não foi capaz de mudar a percepção do entrevistado;
- As palavras por *Cluster* representam bem o que aconteceu no setor;
- O método empregado possui estrutura para responder perguntas gerais e, se parametrizado, possui capacidade de responder perguntas mais específicas, como: O ano de 2024 será o ano de quê? Quais serão os tópicos-chaves de um segmento no próximo ano?;
- O método também possui capacidade para medir se um assunto é classificado como morto ou na moda, pois as pessoas falam daquilo que têm interesse, logo, quanto maior a repetição, maior o foco;
- Essa lógica pode ser aplicada no âmbito político ou pode fomentar um produto, via monitoramento de mídia;
- Com a parte de Análise de sentimento, é possível fazer correções de estratégia e de tática, e ela cria um viés para ajudar a ler as notícias;
- A metodologia permite um entendimento novo do setor, a geração de um conhecimento prévio, no qual é possível perceber se o modelo mental estava preso a um viés, ou não;
- O método permite o teste de uma hipótese ou de uma premissa;

De forma geral, é possível inferir que o método é capaz de ajudar no entendimento do setor, aumentando a capacidade de compreensão e de percepção do setor. Porém os dados coletados podem limitar essa capacidade mais ampla caso tratem, na sua maioria, de um mesmo assunto. Logo, os resultados obtidos podem ser assuntos habituais do setor. Isso explica, em parte, o fato de os *Clusters* estarem sobrepostos, conforme

demonstrado anteriormente. Essa sobreposição pode ter dificultado a interpretação dos *Clusters* ou ter exigido mais esforço de compreensão, afetando, consequentemente, o entendimento das palavras. Isso indica a necessidade de um reprocessamento dos *Clusters*, um refinamento, no sentido de retirar as palavras mais representativas, removendo a contaminação do óbvio e especificando ainda mais as narrativas para direcionar a análise.

Como a ferramenta foi desenvolvida para atender um grupo mais abrangente de analistas setoriais, o foco ficou restrito ao setor de mineração de minerais, proporcionando à ferramenta fornecer mais contexto e possibilitando uma análise de mercado mais profunda. Mesmo fornecendo um norte, dando uma direção para onde e para o que olhar, é necessário um aprofundamento. Segundo os entrevistados, esse aprofundamento vem da visão de negócio, da necessidade mais focada do uso da informação.

Relativo à questão 11, os entrevistados relataram que perceberam/identificaram as temáticas de “Desempenho da economia”, de “Outros” e de “Novas tecnologias”, totalizando 68% das citações. Um dos entrevistados observou que os dados estão focados em bolsa de valores, o que justifica o fato da temática “Desempenho da economia” ter sido a mais citada. Além disso, a maior parte dos respondentes trabalha no ecossistema de *Big Data* e em suas inúmeras vertentes, o que justifica a relevância da temática “Novas tecnologias”.

Tabela 7 – Temas de Monitoramento Citados Percebidos/Identificados

Temas de Monitoramento percebidas/identificadas	Frequência de Citação	Part. %
Desempenho da economia	9	29,03%
Outros	7	22,58%
Novas tecnologias	5	16,13%
Aspectos culturais	2	6,45%
Mudança de estratégia	2	6,45%
Mudanças na legislação ou normas regulatórias	2	6,45%
Parcerias, fusões e aquisições	2	6,45%
Questões ligadas ao meio ambiente e sociais	2	6,45%
Total Geral	31	100,00%

Fonte: elaborado pelo autor.

De forma geral, de acordo com os relatos dos entrevistados, a metodologia possui muita capacidade de criar sentido e de entender um setor. Quanto à sua capacidade de mudar a compreensão, ou à percepção do setor, seja do todo, seja de apenas uma parte, essa capacidade pode ser entendida como média-alta, devido aos 7 respondentes que disseram média capacidade e aos 7 que disseram muita capacidade. Quanto ao

monitoramento dos aspectos ambientais, os entrevistados perceberam que a metodologia é capaz de acompanhar mudanças em termos de economia e de tecnologia. Essa característica pode ser influenciada pelos dados coletados.

Como o propósito central é desenvolver uma metodologia para monitorar a evolução de mudanças estruturais no macroambiente setorial, no caso dessa pesquisa, o setor de mineração de minerais no Brasil, pode-se deduzir que o método empregado possui capacidade de monitorar a evolução de mudanças estruturais no macroambiente do setor de mineração no Brasil, de criar sentido e de permitir uma compreensão maior e melhor desse setor, além de possibilitar mudanças em sua compreensão.

6 CONCLUSÃO E CONSIDERAÇÕES FINAIS

O objetivo dessa pesquisa foi desenvolver uma metodologia de recuperação da informação a partir da aplicação de sumarização automática de texto para monitorar a evolução de mudanças estruturais no macroambiente do setor de mineração no Brasil. A sumarização de texto é uma etapa importante na análise de inúmeros blocos de textos desestruturados para obter *insights* e recuperar informações relevantes para as organizações. A aplicação da sumarização de texto nessa pesquisa permitiu a redução da quantidade de em caracteres~~,~~ cerca de 75%, possibilitando a inserção de outras técnicas de NLP.

Uma delas foi a mensuração da similaridade textual, que permitiu identificar quais notícias narram a mesma história, evitando, dessa forma, a repetição de termos que pudessem enviesar a análise ou mascarar um resultado mais importante. Outra técnica de NLP aplicada foi a classificação de sentimento, sendo ele positivo, neutro ou negativo. Cada notícia recebeu uma classificação, que permitiu compreender qual a linha editorial desenvolvida sobre o setor de mineração. Notícias sobre acidentes impactam essa classificação.

Para todas as técnicas de NLP citadas até aqui, o *transformer* BERT foi utilizado, devido a sua versatilidade.

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 de todo esse ferramental permitiu identificar que há um efeito longitudinal, ou seja, a data da notícia deve ser levada em consideração, pois influencia na análise. Quanto mais agregados os dados estiverem na análise, maior será esse efeito, tendo em vista que um *Cluster* pode conter mais notícias de um determinado ano ou mês, aumentando a repetição de algumas palavras. Essas palavras com maior frequência evidenciam a sua importância dentro do discurso e podem formar um *Cluster* concentrado nessas palavras mais relevantes de uma determinada data. Quanto a isso, é necessária cautela no momento da análise, pois a repetição das notícias pode ser potencializada para atender algum fim específico, ou simplesmente demonstrar o grau de importância daquele assunto naquele momento do tempo.

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

Dadas as características das notícias, há uma tendência de alta sobreposição dos *Clusters*, tanto de forma geral como por ano. Os valores da pontuação *Silhouette* tendem a 0, na maioria dos casos, não fornecendo para o algoritmo critérios claros e precisos para a classificação das notícias. Os dados mostraram várias notícias na borda de decisão, de forma que elas poderiam ser classificadas em um *Cluster* ou em outro. Devido a isso, alguns *Clusters* possuíam um núcleo denso e algumas notícias se dissipavam em torno desse núcleo, evidenciando o compartilhamento de palavras com alguma variação no discurso. As notícias que se afastavam desse núcleo tratavam de um assunto diferente do *core* do núcleo do *Cluster*.

Diante desse fato, é possível concluir que a maioria das notícias utiliza o mesmo agrupamento de palavras. Esse agrupamento de palavras ocorre devido à sua repetição e por elas estarem em dois ou três *Clusters*, justificando, em parte, o alto grau de sobreposição entre os *Clusters*. Isso indica que há um padrão no discurso, sendo que a maioria das notícias possuem a mesma narrativa. Um dos achados é a influência das fontes das notícias, pois a maioria coletada para essa pesquisa é oriunda de fontes que falam do comportamento das empresas em bolsa de valores. Esse padrão de notícias trata basicamente da produção no trimestre e no ano, principalmente da empresa Vale, por ser a maior mineradora do país, e da oscilação da bolsa de valores de São Paulo, a IBOVESPA, hoje chamada de B3.

Outros fatos observados na análise foram o surgimento de empresas, as notícias relacionadas a projeto de ouro, a repercussão do projeto de mineração em terras indígenas, as questões ligadas à energia, os projetos de *startups* focadas em mineração, o problema das barragens da empresa Vale em Minas Gerais e as movimentações estratégicas das empresas. O método empregado também se mostrou eficaz em separar notícias que não estavam relacionadas à mineração de minerais. Algumas notícias relacionadas à mineração de criptomoedas e à mineração de texto foram captadas de forma não intencional, mas, mesmo assim, foram agrupadas em *Clusters* específicos, não contaminando a análise do setor de mineração. O *transformer* BERT também se mostrou eficaz em processar notícias no idioma inglês, e o algoritmo de *Clustering* foi eficiente ao não misturar notícias do idioma inglês com o português.

Os resultados da Análise de Sentimento, tanto por contexto quanto por data, evidenciam a repercussão dos fatos de mineração de minerais e como eles foram assimilados. Há uma tendência de notícias negativas serem mais replicadas do que as positivas, e como notícias sobre mineração de minérios tendem a ser assimiladas como negativas, o setor, como um todo, deve gerir ou acompanhar o que falam sobre ele nas mídias sociais, monitorando os aspectos cultural e social.

A metodologia, quando plotada no *Dashboard*, evidenciou a captação de sinais fracos quanto a alguns temas. Isso ficou explicitado quando alguns *Clusters* trataram de narrativa/discursos específicos em uma determinada data. A ferramenta também captou os principais *drivers* a serem monitorados. Esses *drivers* são as palavras mais representativas por *Cluster* e as narrativas/discursos, tanto quanto a frequência, de forma constante ou pontual, ou de um tema, devem ser monitoradas a partir de certo momento. O monitoramento dos *drivers* fornece uma direção aos gerentes sobre quais temas eles devem focar quando se trata de pontos-chaves de monitoramento de inteligência.

A metodologia desenvolvida evidencia sua capacidade de ser aplicada junto a técnicas envolvidas na análise de negócios, a dados competitivos e a informações, tanto as clássicas quanto as contemporâneas, de inteligência e de estratégia competitiva. Essas técnicas de análise de negócios, dados competitivos e informações demandam um grande volume textual para a geração de sentido, principalmente em cada parte dos *frames* de cada técnica. Elas também ajudam no ordenamento da informação, possibilitando a análise e, consequentemente, a derivação de conclusões.

Por fim, a metodologia e o processo desenvolvido mostraram ser capazes de processar multidocumentos em um domínio específico para gerar uma saída sumarizada, aplicada ao setor de mineração de minerais. Nos anos com maior quantidade de notícias, os dados mostram uma maior repetição semântica, classificada como negativa em termo de sentimento, com um discurso relacionado à variação da bolsa de São Paulo e à produção de minério de ferro e ouro ao longo dos anos. Fora essa temática, cada ano abordou temas pertinentes à mineração.

A partir da teoria e dos resultados empíricos analisados anteriormente, podemos enumerar uma série de caminhos de pesquisa que poderiam resultar em melhorias metodológicas, detalhados a seguir:

1. Considerar retreinar o BERT para a tarefa de Análise de Sentimento no idioma português e focado no setor de mineração para uma análise de sentimento mais assertiva, pois o BERT permite esse tipo de parametrização, de idioma e de domínio, para obter um resultado melhor;
2. Criar um processo de classificação de texto baseado nas características dos *Clusters* para identificar qual o assunto do texto e para o significado do texto ser classificado em um *Cluster*, e não em outro;
3. Parametrizar a ferramenta para indicar quais os principais assuntos do setor dentro de um intervalo de tempo;
4. Responder perguntas mais específicas, como, por exemplo, qual a tendência do preço do minério de ferro;

5. Reprocessar os dados para retirar redundâncias, por meio da extração das palavras mais representativas do processo, e, assim, melhorar o resultado, identificando questões até então não identificadas;
6. Criar dois *Dashboards* para consumo da informação, sendo um com repetição semântica, para captar temas considerados de maior interesse, e outro sem repetição semântica, para captar temas e assuntos pontuais e importantes;
7. Reconstruir todo o processo de *Clustering* utilizando *bi-gram* para captar um entendimento melhor do que está sendo falado e comparar com o método utilizado para avaliar qual entrega um resultado de compreensão mais claro;
8. Refazer todo o processo metodológico utilizando o *Chat-GPT* e suas variações ao invés do BERT;
9. Inserir notas explicativas ao longo do *Dashboard* para ajudar na interpretação dos resultados;
10. Expandir o modelo com o uso de dicionários, tesauros e ontologias, trazendo mais sentido e significado, tanto humano quanto para a máquina.

Como foi apontado na introdução desta tese, a grande quantidade de textos cria uma sobrecarga de informação, sendo necessários métodos cada vez mais eficientes de recuperação da informação focada na necessidade do demandante. Tendo isso como motivação, essa tese se propôs a desenvolver um método que apontasse um caminho para esse objetivo dentro de um contexto setorial específico, utilizando a sumarização de texto como etapa inicial. Com base nas modernas arquiteturas de *Deep Learning* para NLP, também chamadas de *Transformers*, foi possível empregar diversas técnicas de NLP usando semântica. Diante disso, acreditamos na maior exploração desse campo no futuro, sustentada pelo avanço da inteligência artificial e pela construção de dicionários para os *transformers*.

REFERÊNCIAS

ABDEL-SALAM, Shehab; RAFEA, Ahmed. Performance Study on Extractive Text Summarization Using BERT Models. **Information**, *[S. l.]*, v. 13, n. 2, p. 67, 2022. DOI: 10.3390/info13020067. Disponível em: https://www.mdpi.com/2078-2489/13/2/67.

AGATHOKLEOUS, Evgenios; RILLIG, Matthias C.; PEÑUELAS, Josep; YU, Zhen. One hundred important questions facing plant science derived using a large language model. **Trends in Plant Science**, *[S. l.]*, v. 29, n. 2, p. 210–218, 2024. DOI: 10.1016/j.tplants.2023.06.008. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1360138523001991.

ALAMI, Nabil; MEKNASSI, Mohammed; EN-NAHNAHI, Noureddine. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. **Expert Systems with Applications**, *[S. l.]*, v. 123, p. 195–211, 2019. DOI: 10.1016/j.eswa.2019.01.037. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417419300375.

ALAMI, Nabil; MEKNASSI, Mohammed; EN-NAHNAHI, Noureddine; EL ADLOUNI, Yassine; AMMOR, Ouafae. Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling. **Expert Systems with Applications**, *[S. l.]*, v. 172, p. 114652, 2021. DOI: 10.1016/j.eswa.2021.114652. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417421000932.

ALLAN, James; GUPTA, Rahul; KHANDELWAL, Vikas. Temporal summaries of new topics. *In*: PROCEEDINGS OF THE 24TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL - SIGIR ’01 2001, New York, New York, USA. **Anais** [...]. New York, New York, USA: ACM Press, 2001. p. 10–18. DOI: 10.1145/383952.383954. Disponível em: http://portal.acm.org/citation.cfm?doid=383952.383954.

ALOM, Md Zahangir et al. A State-of-the-Art Survey on Deep Learning Theory and Architectures. **Electronics**, *[S. l.]*, v. 8, n. 3, p. 292, 2019. DOI: 10.3390/electronics8030292. Disponível em: https://www.mdpi.com/2079-9292/8/3/292.

AVERSA, Joseph; HERNANDEZ, Tony; DOHERTY, Sean. Incorporating big data within retail organizations: A case study approach. **Journal of Retailing and Consumer Services**, *[S. l.]*, v. 60, n. December 2020, p. 102447, 2021. DOI: 10.1016/j.jretconser.2021.102447. Disponível em: https://doi.org/10.1016/j.jretconser.2021.102447.

BALDUINI, Marco; BRAMBILLA, Marco; DELLA VALLE, Emanuele; MARAZZI, Christian; ARABGHALIZI, Tahereh; RAHDARI, Behnam; VESCOVI, Michele. Models

and Practices in Urban Data Science at Scale. **Big Data Research**, *[S. l.]*, v. 17, p. 66–84, 2019. DOI: 10.1016/j.bdr.2018.04.003. Disponível em: https://doi.org/10.1016/j.bdr.2018.04.003.

BONDIELLI, Alessandro; MARCELLONI, Francesco. On the use of summarization and transformer architectures for profiling résumés. **Expert Systems with Applications**, *[S. l.]*, v. 184, p. 115521, 2021. DOI: 10.1016/j.eswa.2021.115521. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417421009301.

BRANDS, Kritine. **Big Data and Business Intelligence for Management Accountants**. [s.l.] : Tech Practices, 2014.

CHOWDHURY, Gobinda G. Natural language processing. **Annual Review of Information Science and Technology**, *[S. l.]*, v. 37, n. 1, p. 51–89, 2003. DOI: https://doi.org/10.1002/aris.1440370103. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103.

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**, *[S. l.]*, v. 7, n. 4, p. 285, 2016. DOI: 10.21512/comtech.v7i4.3746. Disponível em: https://journal.binus.ac.id/index.php/comtech/article/view/3746.

COLLOBERT, Ronan; WESTON, Jason; BOTTOU, Léon; KARLEN, Michael; KAVUKCUOGLU, Koray; KUKSA, Pavel. Natural Language Processing (Almost) from Scratch. **J. Mach. Learn. Res.**, *[S. l.]*, v. 999888, p. 2493–2537, 2011. Disponível em: http://dl.acm.org/citation.cfm?id=2078183.2078186.

COMAI, Alessandro; MILLÁN, Joaquín. **Mapping & anticipating the competitive landscape**. [s.l.] : Editora Miniera SL, 2006.

COPELAND, MICHAEL. **What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?** 2016. Disponível em: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/. Acesso em: 15 out. 2021.

CRESWELL, John W. **Research design: qualitative, quantitative, and mixed methods approaches**. 4. ed. Los Angeles.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *[S. l.]*, 2018. Disponível em: http://arxiv.org/abs/1810.04805.

DHAR, Vasant. Data science and prediction. **Communications of the ACM**,

[S. l.], v. 56, n. 12, p. 64–73, 2013. DOI: 10.1145/2500499. Disponível em: https://dl.acm.org/doi/10.1145/2500499.

FLEISHER, Craig S.; BENSOUSSAN, Babette E. **Strategic and competitive analysis: methods and tecnhiques for analyzing business competition**. Upper Saddle River: Prentice-Hall, Inc., 2003.

FOSSO WAMBA, Samuel; AKTER, Shahriar; EDWARDS, Andrew; CHOPIN, Geoffrey; GNANZOU, Denis. How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**, *[S. l.]*, v. 165, p. 234–246, 2015. DOI: 10.1016/j.ijpe.2014.12.031.

FURNER, Jonathan. Information Science Is Neither. **Library Trends**, *[S. l.]*, v. 63, n. 3, p. 362–377, 2015. DOI: 10.1353/lib.2015.0009. Disponível em: https://muse.jhu.edu/content/crossref/journals/library_trends/v063/63.3.furner.html.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo.

GOLDSMITH, Jeff; SUN, Yifei; FRIED, Linda P.; WING, Jeannette; MILLER, Gary W.; BERHANE, Kiros. The Emergence and Future of Public Health Data Science. **Public Health Reviews**, *[S. l.]*, v. 42, 2021. DOI: 10.3389/phrs.2021.1604023. Disponível em: https://www.ssph-journal.org/articles/10.3389/phrs.2021.1604023/full.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. [s.l.] : MIT Press, 2016. Disponível em: http://www.deeplearningbook.org.

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**, *[S. l.]*, v. 115, p. 264–275, 2019. DOI: 10.1016/j.eswa.2018.07.047. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417418304743.

GOULARTE, Fábio Bif; WILGES, Beatriz; NASSAR, Silvia Modesto; CISLAGHI, Renato. Métricas de sumarização automática de texto em tarefas de um Ambiente Virtual de Aprendizagem. *In*: 2014, **Anais** [...]. [s.l: s.n.] p. 752. DOI: 10.5753/cbie.sbie.2014.752. Disponível em: http://br-ie.org/pub/index.php/sbie/article/view/3007.

GUIMARÃES, Nuno; CAMPOS, Ricardo; JORGE, Alípio. Pre-trained language models: What do they know? **WIREs Data Mining and Knowledge Discovery**, *[S. l.]*, v. 14, n. 1, 2024. DOI: 10.1002/widm.1518. Disponível em: https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1518.

HAMET, Pavel; TREMBLAY, Johanne. Artificial intelligence in medicine. **Metabolism**, *[S. l.]*, v. 69, p. S36–S40, 2017. DOI: 10.1016/j.metabol.2017.01.011. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S002604951730015X.

HARK, Cengiz; KARCI, Ali. Karcı summarization: A simple and effective approach for automatic text summarization using Karcı entropy. **Information Processing & Management**, *[S. l.]*, v. 57, n. 3, p. 102187, 2020. DOI: 10.1016/j.ipm.2019.102187. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0306457319306867.

HUANG, Allen H.; WANG, Hui; YANG, Yi. <scp>FinBERT</scp> : A Large Language Model for Extracting Information from Financial Text*. **Contemporary Accounting Research**, *[S. l.]*, v. 40, n. 2, p. 806–841, 2023. DOI: 10.1111/1911-3846.12832. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/1911-3846.12832.

IBRAM. **Mineração em Números - 4º Trimestre de 2020**. [s.l: s.n.]. Disponível em: https://ibram.org.br/wp-content/uploads/2021/02/Infografico-Mineracao-em-Numeros-2020-NOVO-1-1.pdf.

IWASAKI, Yuuki; YAMASHITA, Akihiro; KONNO, Yoko; MATSUBAYASHI, Katsushi. Japanese Abstractive Text Summarization using BERT. **Advances in Science, Technology and Engineering Systems Journal**, *[S. l.]*, v. 5, n. 6, p. 1674–1682, 2020. DOI: 10.25046/aj0506199. Disponível em: https://astesj.com/v05/i06/p199/.

JAIN, Priyank; GYANCHANDANI, Manasi; KHARE, Nilay. Big data privacy: a technological perspective and review. **Journal of Big Data**, *[S. l.]*, v. 3, n. 1, p. 25, 2016. DOI: 10.1186/s40537-016-0059-y. Disponível em: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0059-y.

JOHN, Ansamma; PREMJITH, P. S.; WILSCY, M. Extractive multi-document summarization using population-based multicriteria optimization. **Expert Systems with Applications**, *[S. l.]*, v. 86, p. 385–397, 2017. DOI: 10.1016/j.eswa.2017.05.075. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417417304049.

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**, *[S. l.]*, v. 129, p. 200–215, 2019. DOI: 10.1016/j.eswa.2019.03.045. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417419302192.

JOSHI, ARAVIND K. Natural Language Processing. **Science**, *[S. l.]*, v. 253, n. 5025, p. 1242–1249, 1991. DOI: 10.1126/science.253.5025.1242. Disponível em: https://science.sciencemag.org/content/253/5025/1242.

KHAMPARIA, Aditya; SINGH, Karan Mehtab. A systematic review on deep learning architectures and applications. **Expert Systems**, *[S. l.]*, v. 36, n. 3, p. e12400, 2019. DOI: 10.1111/exsy.12400. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/exsy.12400.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 5. ed. São Paulo.

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**, *[S. l.]*, v. 167, p. 114152, 2021. a. DOI: 10.1016/j.eswa.2020.114152. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417420308952.

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**, *[S. l.]*, p. 016555152199061, 2021. b. DOI: 10.1177/0165551521990616. Disponível em: http://journals.sagepub.com/doi/10.1177/0165551521990616.

LANCASTER, Frederick Wilfrid. **Indexação e Resumos - Teoria e Pratica**. 2ª edição ed. Brasília.

LEIJNEN, Stefan; VEEN, Fjodor Van. The Neural Network Zoo. **Proceedings**, *[S. l.]*, v. 47, n. 1, p. 9, 2020. DOI: 10.3390/proceedings47010009. Disponível em: https://www.mdpi.com/2504-3900/47/1/9.

LI, Ping; YU, Jiong. Extractive Summarization Based on Dynamic Memory Network. **Symmetry**, *[S. l.]*, v. 13, n. 4, p. 600, 2021. DOI: 10.3390/sym13040600. Disponível em: https://www.mdpi.com/2073-8994/13/4/600.

LIU, Yang; LAPATA, Mirella. Text summarization with pretrained encoders. **EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference**, *[S. l.]*, p. 3730–3740, 2019. DOI: 10.18653/v1/d19-1387.

LIU, Yixin; SHI, Kejian; HE, Katherine S.; YE, Longtian; FABBRI, Alexander R.; LIU, Pengfei; RADEV, Dragomir; COHAN, Arman. On Learning to Summarize with Large Language Models as References. *[S. l.]*, 2023. Disponível em: http://arxiv.org/abs/2305.14239.

LUHN, H. P. The Automatic Creation of Literature Abstracts. **IBM Journal of Research and Development**, *[S. l.]*, v. 2, n. 2, p. 159–165, 1958. DOI:

10.1147/rd.22.0159. Disponível em: http://ieeexplore.ieee.org/document/5392672/.

MA, Congbo; ZHANG, Wei Emma; GUO, Mingyu; WANG, Hu; SHENG, Quan Z. Multi-document Summarization via Deep Learning Techniques: A Survey. *[S. l.]*, 2020. Disponível em: http://arxiv.org/abs/2011.04843.

MAEDA, Takashi. An approach toward functional text structure analysis of scientific and technical documents. **Information Processing & Management**, *[S. l.]*, v. 17, n. 6, p. 329–339, 1981. DOI: 10.1016/0306-4573(81)90047-9. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/0306457381900479.

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

MILLER, Derek. Leveraging BERT for Extractive Text Summarization on Lectures. *[S. l.]*, 2019. Disponível em: http://arxiv.org/abs/1906.04165.

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

MOENS, Marie-Francine; DUMORTIER, Jos. Use of a text grammar for generating highlight abstracts of magazine articles. **Journal of Documentation**, *[S. l.]*, v. 56, n. 5, p. 520–539, 2000. DOI: 10.1108/EUM0000000007126. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/EUM0000000007126/full/html.

MORADI, Milad; DASHTI, Maedeh; SAMWALD, Matthias. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. **Journal of Biomedical Informatics**, *[S. l.]*, v. 107, p. 103452, 2020. DOI: 10.1016/j.jbi.2020.103452. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1532046420300800.

MORADI, Milad; DORFFNER, Georg; SAMWALD, Matthias. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. **Computer Methods and Programs in Biomedicine**, *[S. l.]*, v. 184, p. 105117, 2020. DOI: 10.1016/j.cmpb.2019.105117. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S016926071931137X.

MUTLU, Begum; SEZER, Ebru A.; AKCAYOL, M. Ali. Candidate sentence selection for extractive text summarization. **Information Processing & Management**, *[S. l.]*, v. 57, n. 6, p. 102359, 2020. DOI: 10.1016/j.ipm.2020.102359. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0306457320308542.

PADMAKUMAR, Aishwarya; SARAN, Akanksha. Unsupervised Text Summarization Using Sentence Embeddings. *[S. l.]*, 2016. Disponível em: https://pdfs.semanticscholar.org/253e/17a1692f0d6351345435a59a80f5ddb17731.pdf.

PORTER, Michael E. **Estratégia Competitiva - Técnicas Para Análise de Indústrias e da Concorrência**. 1a. ed. [s.l.] : GEN Atlas, 2005.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar. **Metodologia do trabalho científico [recurso eletrônico] : métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. 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**, *[S. l.]*, v. 10, n. 12, p. 44–53, 2018. DOI: 10.5815/ijmecs.2018.12.06. Disponível em: http://www.mecs-press.org/ijmecs/ijmecs-v10-n12/v10n12-6.html.

RAMEZANI, Majid; FEIZI-DERAKHSHI, Mohammad-Reza. AUTOMATED TEXT SUMMARIZATION: AN OVERVIEW. **Applied Artificial Intelligence**, *[S. l.]*, v. 28, n. 2, p. 178–215, 2014. DOI: 10.1080/08839514.2014.862783. Disponível em: http://www.tandfonline.com/doi/abs/10.1080/08839514.2014.862783.

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**, *[S. l.]*, v. 38, n. 2, p. 56–68, 2009. DOI: 10.1590/s0100-19652009000200005.

RICHARDSON, Roberto Jarry. **Pesquisa Social: Métodos e Técnicas**. 3. ed 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**, *[S. l.]*, v. 169, n. July 2020, 2021. DOI: 10.1016/j.eswa.2020.114320.

RUSH, J. E.; SALVADOR, R.; ZAMORA, A. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. **Journal of the American Society for Information Science**, *[S. l.]*, v. 22, n. 4, p. 260–274, 1971. DOI: 10.1002/asi.4630220405. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.4630220405.

SALEHI, Hadi; BURGUEÑO, Rigoberto. Emerging artificial intelligence methods in structural engineering. **Engineering Structures**, *[S. l.]*, v. 171, p. 170–189, 2018. DOI: 10.1016/j.engstruct.2018.05.084. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0141029617335526.

SARKAR, Kamal. Using Domain Knowledge for Text Summarization in Medical Domain. **International Journal of Recent Trends in Engineering**, *[S. l.]*, v. 1, 2009.

SEARLE, Thomas; IBRAHIM, Zina; TEO, James; DOBSON, Richard JB. Estimating Redundancy in Clinical Text. *[S. l.]*, 2021. Disponível em: http://arxiv.org/abs/2105.11832.

SENGUPTA, Saptarshi; BASAK, Sanchita; SAIKIA, Pallabi; PAUL, Sayak; TSALAVOUTIS, Vasilios; ATIAH, Frederick; RAVI, Vadlamani; PETERS, Alan. A review of deep learning with special emphasis on architectures, applications and recent trends. **Knowledge-Based Systems**, *[S. l.]*, v. 194, p. 105596, 2020. DOI: 10.1016/j.knosys.2020.105596. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S095070512030071X.

SHI, Yiwen et al. Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting. **Journal of Biomedical Informatics**, *[S. l.]*, v. 148, p. 104533, 2023. DOI: 10.1016/j.jbi.2023.104533. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S153204642300254X.

SHRESTHA, Ajay; MAHMOOD, Ausif. Review of Deep Learning Algorithms and Architectures. **IEEE Access**, *[S. l.]*, v. 7, p. 53040–53065, 2019. DOI: 10.1109/ACCESS.2019.2912200. Disponível em: https://ieeexplore.ieee.org/document/8694781/.

SINHA, Aakash; YADAV, Abhishek; GAHLOT, Akshay. Extractive Text Summarization using Neural Networks. *[S. l.]*, 2018. Disponível em: http://arxiv.org/abs/1802.10137.

SNIR, Marc. Computer and information science and engineering. **Communications of the ACM**, *[S. l.]*, v. 54, n. 3, p. 38–43, 2011. DOI: 10.1145/1897852.1897867. Disponível em: https://dl.acm.org/doi/10.1145/1897852.1897867.

SONG, Shengli; HUANG, Haitao; RUAN, Tongxiao. Abstractive text summarization using LSTM-CNN based deep learning. **Multimedia Tools and Applications**, *[S. l.]*, v. 78, n. 1, p. 857–875, 2019. DOI: 10.1007/s11042-018-5749-3. Disponível em: http://link.springer.com/10.1007/s11042-018-5749-3.

SOUZA, Renato Rocha. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. Belo HorizonteEscola de Ciência da Informação - UFMG, , 2005.

SRIKANTH, Anirudh; UMASANKAR, Ashwin Shankar; THANU, Saravanan; NIRMALA, S. Jaya. Extractive Text Summarization using Dynamic Clustering and Co-Reference on BERT. *In*: 2020 5TH INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND SECURITY (ICCCS) 2020, **Anais** [...]. : IEEE, 2020. p. 1–5. DOI: 10.1109/ICCCS49678.2020.9277220. Disponível em: https://ieeexplore.ieee.org/document/9277220/.

STANTON, Jeffrey M. **Data science: what's in it for the new librarian?** 2012. Disponível em: https://ischool.syr.edu/infospace/2012/07/16/data-science-whats-in-it-for-the-new-librarian/. Acesso em: 19 jun. 2020.

SYED, Ayesha Ayub; GAOL, Ford Lumban; MATSUO, Tokuro. A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. **IEEE Access**, *[S. l.]*, v. 9, p. 13248–13265, 2021. DOI: 10.1109/ACCESS.2021.3052783. Disponível em: https://ieeexplore.ieee.org/document/9328413/.

TAN, Bowen; KIEUVONGNGAM, Virapat; NIU, Yiming. Automatic text summarization of COVID-19 medical research articles using BERT and GPT-2. **arXiv**, *[S. l.]*, 2020.

TAYAL, Madhuri A.; RAGHUWANSHI, Mukesh M.; MALIK, Latesh G. ATSSC: Development of an approach based on soft computing for text summarization. **Computer Speech & Language**, *[S. l.]*, v. 41, p. 214–235, 2017. DOI: 10.1016/j.csl.2016.07.002. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S088523081630208X.

VAN VEEN, Dave et al. Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. *[S. l.]*, 2023. Disponível em: http://arxiv.org/abs/2309.07430.

VASCONCELLOS, Vera Maria Ramos De; NASCIMENTO DA SILVA, Anne Patrícia Pimentel; DE SOUZA, Roberta Teixeira. O Estado da Arte ou o Estado do Conhecimento. **Educação**, *[S. l.]*, v. 43, n. 3 SE-Outros Temas, p. e37452, 2020. DOI: 10.15448/1981-2582.2020.3.37452. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/faced/article/view/37452.

WANG, Lin. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, *[S. l.]*, v. 74, n. 6, p. 1243–1257, 2018. DOI: 10.1108/JD-02-2018-0036.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. 2. ed. Ro de Janeiro.

WEAVER, Adam. Tourism, big data, and a crisis of analysis. **Annals of Tourism Research**, *[S. l.]*, v. 88, p. 103158, 2021. DOI: 10.1016/j.annals.2021.103158. Disponível em: https://doi.org/10.1016/j.annals.2021.103158.

WEISSENBERGER, Lynnsey. Toward a universal, meta-theoretical framework for music information classification and retrieval. **Journal of Documentation**, *[S. l.]*, v. 71, n. 5, p. 917–937, 2015. DOI: 10.1108/JD-08-2013-0106. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-08-2013-0106/full/html.

WILKS, Yorick. The History of Natural Language Processing and Machine Translation. *In*: [s.l: s.n.]. p. 14.

WOLF, Thomas et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *[S. l.]*, 2019. Disponível em: http://arxiv.org/abs/1910.03771.

WU, Ning; GONG, Ming; SHOU, Linjun; LIANG, Shining; JIANG, Daxin. Large Language Models are Diverse Role-Players for Summarization Evaluation. *[S. l.]*, 2023. Disponível em: http://arxiv.org/abs/2303.15078.

XIE, Qianqian; LUO, Zheheng; WANG, Benyou; ANANIADOU, Sophia. A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models. *[S. l.]*, 2023. Disponível em: http://arxiv.org/abs/2304.08763.

YANG, Guangbing; CHEN, Nian-Shing; KINSHUK; SUTINEN, Erkki; ANDERSON, Terry; WEN, Dunwei. The effectiveness of automatic text summarization in mobile learning contexts. **Computers & Education**, *[S. l.]*, v. 68, p. 233–243, 2013. DOI: 10.1016/j.compedu.2013.05.012. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0360131513001334.

YANG, Min; WANG, Xintong; LU, Yao; LV, Jianming; SHEN, Ying; LI, Chengming. Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint. **Information Sciences**, *[S. l.]*, v. 521, p. 46–61, 2020. DOI: 10.1016/j.ins.2020.02.040. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0020025520301225.

YANG, Xianjun; LI, Yan; ZHANG, Xinlu; CHEN, Haifeng; CHENG, Wei. Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization. *[S. l.]*, 2023. Disponível em: http://arxiv.org/abs/2302.08081.

YOU, Fucheng; ZHAO, Shuai; CHEN, Jingjing. A Topic Information Fusion and Semantic Relevance for Text Summarization. **IEEE Access**, *[S. l.]*, v. 8, p. 178946–178953, 2020. DOI: 10.1109/ACCESS.2020.2999665. Disponível em: https://ieeexplore.ieee.org/document/9107114/.

ZHANG, Aston; LIPTON, Zachary C.; LI, Mu; SMOLA, Alexander J. **Dive into Deep Learning**. [s.l: s.n.]. Disponível em: https://d2l.ai.

ZHANG, Haopeng; LIU, Xiao; ZHANG, Jiawei. SummIt: Iterative Text Summarization via ChatGPT. *[S. l.]*, 2023. Disponível em: http://arxiv.org/abs/2305.14835.