UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA, GESTÃO E MÍDIA DO CONHECIMENTO

Letícia Silveira Artese

MODELO DE DESCOBERTA DE CONHECIMENTO EM TEXTO PARA DETECÇÃO DE SINAIS FRACOS PARA TECNOLOGIAS EMERGENTES

Florianópolis
2023

Letícia Silveira Artese

MODELO DE DESCOBERTA DE CONHECIMENTO EM TEXTO PARA DETECÇÃO DE SINAIS FRACOS PARA TECNOLOGIAS EMERGENTES

Tese submetida ao Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Doutora em Engenharia do Conhecimento.

Orientador: Prof. Alexandre Leopoldo Gonçalves, Dr.
Coorientador: Prof. José Leomar Todesco, Dr.
Coorientador externo: Prof. José Millet Roig, PhD.

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Artese, Leticia Silveira
Modelo de Descoberta de Conhecimento em Texto para Detecção de Sinais Fracos para Tecnologias Emergentes / Leticia Silveira Artese ; orientador, Alexandre Leopoldo Gonçalves, coorientador, José Leomar Todesco, coorientador, José Millet Roig, 2023.
223 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2023.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Mineração de Texto. 3. Sinais Fracos. 4. Foresight. 5. Inteligência Antecipativa. I. Gonçalves, Alexandre Leopoldo. II. Todesco, José Leomar. III. Roig, José Millet IV. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Engenharia e Gestão do Conhecimento. V. Título.

Letícia Silveira Artese

MODELO DE DESCOBERTA DE CONHECIMENTO EM TEXTO PARA DETECÇÃO DE SINAIS FRACOS PARA TECNOLOGIAS EMERGENTES

O presente trabalho em nível de Doutorado foi avaliado e aprovado, em 17 de agosto de 2023, pela banca examinadora composta pelos seguintes membros:

Profa. Gertrudes Aparecida Dandolini, Dra.
Instituição Universidade Federal de Santa Catarina

Prof. Denilson Sell, Dr.
Instituição Universidade Federal de Santa Catarina

Prof. Alexandre Moraes Ramos, Dr.
Instituição Universidade Federal de Santa Catarina

Profa. Ana Estela Antunes da Silva, Dra.
Instituição Universidade Estadual de Campinas

Prof. Israel Griol-Barres, Ph.D (membro extraoficial)
Instituição Universitat Politècnica de València

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutora em Engenharia do Conhecimento.

Documento assinado digitalmente
Roberto Carlos dos Santos Pacheco
Data: 06/10/2023 12:38:20-0300
CPF: ***.543.830-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Programa de Pós-Graduação

Documento assinado digitalmente
ALEXANDRE LEOPOLDO GONCALVES
Data: 06/10/2023 11:55:52-0300
CPF: ***.740.349-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Alexandre Leopoldo Gonçalves, Dr.
Orientador

Florianópolis, 2023

AGRADECIMENTOS

Para o resultado desses anos de pesquisa se concluirem nesta tese muito empenho foi necessário. Dedicação não só minha, e esse pequeno espaço é dedicado ao reconhecimento da imensa importância de tantas outras pessoas envolvidas neste processo.

Gostaria de iniciar agradecendo ao meu orientador, professor Alexandre Leopoldo Gonçalves, sem o qual esse trabalho não aconteceria. Para aqueles que não tiveram o prazer de conhecê-lo, me permitam tentar expressar um pouco de quem ele é para mim. Muito além de professor de engenharia de computação e engenharia do conhecimento, é professor da vida. É exemplo de integridade, comprometimento e dedicação. É exemplo de humanidade, acolhedor, empático e solícito. Tantas qualidades infelizmente muitas vezes difíceis de encontrar no meio acadêmico. Já é com saudades que escrevo sobre a falta que sentirei de nossas conversas semanais. Obrigada por confiar em minhas ideias, pelo incentivo e acreditar em minha capacidade. Agradeço pelos momentos de compreensão, pelo apoio intelectual e emocional. Você foi luz e inspiração no meu caminho. Desejo muito sucesso em sua trajetória.

Na sequência gostaria de mencionar meu profundo e sincero agradecimento ao meu namorido, Samuel. Que me ama permitindo ser quem sou, mesmo não gostando das distâncias, está sempre me apoiando nas aventuras que invento. Me mostrou o que é compromisso e que relação se constrói com paciência e dedicação. Amo você! Ansiosa pelo que ainda iremos construir juntos.

Aos meus pais, pela base e apoio me dando coragem para que eu criasse meu caminho e astúcia para percorrê-lo.

Agradeço ao PPGEGC e ao professor Alexandre Biz pela oportunidade de experiência internacional. Agradeço imensamente os professores Pepe e Israel pelo caloroso acolhimento na UPV. E com aperto no coração agradeço meus amigos valencianos: Guille, Milena, Maggie, Jaume, Santi, Irene, ixi tanta gente! Nunca imaginaria ter feito amizade com um grupo de pessoas tão incríveis! *Los echo de menos*!

Um obrigada especial para meu amigo Rato pelas conversas acadêmicas, às vezes nem tanto, revisor de texto, consultor de genealogia e terapeuta nas horas vagas. Agradeço por nunca abandonar sua missão de anos de me fazer acreditar que sou capaz! Por fim, agradeço aos amigos que ficaram em outras cidades, aos amigos que fiz aqui em Florianópolis, aos amigos que reencontrei e ficarão pra sempre (oi Mari!). Aos colegas de sala e projetos do EGC. Obrigada pela compreensão da minha ausência em seus aniversários, noivados, chás de bebê e *open houses*. Fica o registro que estive vibrando por todas as conquistas de vocês à distância.

Obrigada CAPES, FAEPEX, UNICAMP, UFSC, IFSC, USP, UEMG todas as instituições de ensino superior e instituições de fomento brasileiras pelas quais passei. Confio e agradeço pela honra de poder contar com esse apoio. Carrego em mim a responsabilidade social de usufruir desses recursos e prover para o desenvolvimento da ciência nacional.

Para Antônio (*in memorian*) siga sendo anjo.

Obrigada por me permitirem por um
pouquinho de coração no documento.

"Todos esses que aí estão
Atravancando meu caminho,
Eles passarão...
Eu passarinho! " – Mário Quintana

RESUMO

Sinais fracos são fragmentos de informação que a princípio podem parecer vagos e desconexos, mas que atuam como indicadores de um evento futuro e potenciais questões emergentes. Em particular, o domínio tecnológico se destaca no interesse em antecipar informações, visto a estreita relação entre avanço tecnológico e vantagem competitiva. Nessa perspectiva, sinais fracos se mostram uma informação relevante para o planejamento estratégico. O contexto atual de desenvolvimento acelerado enaltece a necessidade de uma abordagem que considere o conceito de incerteza nos planejamentos, viabilizando a antecipação da informação a fim de aproveitar vantagens e minimizar riscos. Diante desse cenário, esta tese propõe um modelo de descoberta de conhecimento em texto para detectar sinais fracos para tecnologias emergentes. A proposta tem sua originalidade marcada pela idealização da identificação de novas palavras como recomendações de sinais fracos. A proposição se sustenta na premissa de compreensão de palavras como o menor nível de abstração de uma informação com base na teoria semiótica de Peirce, e assim, uma forma de buscar a máxima antecipação. O modelo foi elaborado a partir da *Design Science Research Methodology* e tem como foco explorar a semântica presente nos dados não estruturados contido nos documentos de publicações científicas e tecnológicas. O modelo tem como cerne três tarefas de mineração de texto: (1) a identificação de sinais novos: descoberta de palavras novas; (2) a identificação de sinais latentes: monitoramento temporal das palavras; e (3) a identificação de sinais fracos: estimativa sobre o potencial de impacto tecnológico. Para sua operacionalização, métodos e técnicas de processamento de linguagem natural e técnicas de análise de redes temporais foram implementadas, como o modelo para *embedding* BERT e a análise *burst* de redes temporais. A demostração conceitual do modelo e sua operacionalização simulada baseada em um cenário de estudo real apresentaram resultados satisfatórios. Evidenciando a capacidade de obtenção de sinais fracos para tecnologias emergentes a partir da identificação e monitoramento temporal de palavras novas, uma vez que, o modelo é capaz de obter a contextualização dinâmica das palavras novas identificadas. Além disso, como contribuição enaltece-se a viabilidade e importância de abordagens quantitativas e computacionais para a detecção de sinais fracos, particularmente, considerando dados não estruturados como os textos. Assim, o modelo desenvolvido demonstrou competência em prover sinais fracos que apontam para novas tecnologias com antecedência colaborando para processos de planejamento estratégico.

Palavras-chave: Inteligência Antecipativa; Foresight; Estudos Futuros.

ABSTRACT

Weak signals are fragments of information that initially may appear vague or disconnected, but they act as indicators of a future event and potential emerging issues. In particular, the technological domain stands out in its interest to anticipate information, given the close relationship between technological advancement and competitive advantage. In this perspective, weak signals are relevant information for strategic planning. The current context of accelerated development emphasizes the need for an approach that considers the concept of uncertainty in strategic planning, enabling the anticipation of information to take advantages and minimize risks. Given this scenario, this thesis proposes a knowledge discovery in text model for detecting weak signals for emerging technologies. The proposition is premised on the understanding of words as the lowest level of abstraction of information based on Peirce's semiotic theory, and thus, a way to aim for maximum anticipation. The developed model followed the Design Science Research Methodology guidelines and focused on exploring the semantics contained in the unstructured part of the data in documents as scientific and technological publications. The model has three text mining tasks as its core: (1) identification of new signals: discovery of new words; (2) identification of latent signals: temporal monitoring of words; and (3) identification of weak signals: estimation of potential technological impact. For its operationalization methods and techniques of natural language processing and temporal network analysis techniques are implemented, such as the BERT embedding model and burst analysis of temporal networks. The conceptual demonstration of the model and its simulated operationalization based on a real scenario showed satisfactory results. Evidencing the ability to obtain weak signals for emerging technologies by identifying and monitoring new words, once the model is capable of obtain the contextualization of the new identified words. Furthermore, as a contribution, the feasibility and importance of quantitative and computational approaches for detecting weak signals are highlighted, specially approaches toward non structured data as texts. Thus, the developed model proves to be capable of providing weak signals that point to new technologies in advance, thus contributing to strategic planning processes.

Keywords: Anticipatory Intelligence; Foresight; Future Studies.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI - *Artificial Intelligence*

ACM – *Association for Computing Machinery*

AOL – *American Online*

ARPANET - *Advanced Research Projects Agency*

BBC - *British Broadcast Corporation*

BERT - *Bidirectional Encoder Representations from Transformers*

BMBF - *Bundesministerium für Bildung und Forschung*

BPE - *Byte-Pair Encoding*

BOW - *Bag of Words*

CAS - *Complex Adaptative System*

CGEE - Centro de Gestão e Estudos Estratégicos

CPU - *Central Processing Unit*

CQP - *Constained Quadratic Programming*

C&T - Ciência e Tecnologia

CT&I - Ciência, Tecnologia e Inovação

DBS - *Density Bursting Subgraph*

DF - *Document Frequency*

DL - *Deep Learning*

DoD - *Degree of Diffusion*

DoV - *Degree of Visibility*

DS - *Design Science*

DSR - *Design Science Research*

DSRM - *Design Science Research Methodology*

EC - Engenharia do Conhecimento

EGC – Enhenharia e Gestão do Conhecimento

ELMo - *Embeddings from Language Model*

EPO - *European Patent Office*

ETD - *Emergent Technology Detection*

EU - *European Union*

EUA - Estados Unidos da América

FTA - *Future-Oriented Technology Analysis*

FUSE - *Foresight and Understanding from Scientific Exposition*

HSPT - *Horizon Scanning Programme Team*

GC – Gestão do Conhecimento

GloVe - *Global Vectors*

GPT – (2)(3) - *Generative Pre-trained Transformer*

GPU - *Graphics processing unit*

GRU - *Gated recurrent unit*

IARPA - *Intelligence Advanced Research Projects Activity*

IA – Inteligência Artificial

ICZN - *International Code of Zoological Nomenclature*

IDF - Inverse *Document Frequency*

ids – identificadores

IE - *Information Extraction*

IEAc - Inteligência Estratégica Antecipativa coletiva

IEEE - *Institute of Electrical and Electronics Engineers*

INPI - Instituto Nacional de Propriedade Intelectual

IP - *Internet Protocol*

IPC - *International Patent Classification*

IR - *Information Retrieval*

ISO - *International Organization for Standardization*

IUPAC - *International Union of Pure and Applied Chemistry*

IUPAP - *International Union of Pure and Applied Physics*

JPO - *Japan Patent Office*

JRC - *Joint Research Centre*

JSON - *JavaScript Object Notation*

KEM - *Keyword emergence map*

KDD - *Knowledge Discovery in Databases*

KDT - *Knowledge Discovery in Text*

KIM - *Keyword issue map*

KISTI - *Korea Institute of Science and Technology Information*

LDA - *Latent Dirichlet allocation*

LLMs – *Large Language Models*

LOF - *Local Outlier Factor*

LPC - *Linear Predictive Coding*

LSA - *Latent Semantic Analysis*

LSI - *Latent Semantic Indexing*

LSTM - *Long Short-Term Memory*

MDS - *Maximum Density Segment*

MIP - *Mixed Integer Programming*

ML - *Machine Learning*

MLM – *Masked Language Modelling*

NN - *Neural Networks*

NER - *Named Entity Recognition*

NEST – *New and Emerging Signals of Trends*

NISTEP - *National Institute of Science and Technology Policy*

NLP - *Natural Language Processing*

NSP – *Next Sentence Prediction*

OECD - *Organisation for Economic Co-Operation and Development*

OHE - *One-hot encoding*

OTA - *Office of Technology Assessment*

OOV - *Out of Vocabulary*

P&D - Pesquisa e Desenvolvimento

PEST - *Politics*; *Economics*; *Society*; *Technology*

PESTEL - *Political, Economics, Societal, Technological, Environmental and Legal*

PoC - *proof-of-concept*

PPGEGC – Programa de pós-graduação em Engenharia, Gestão e Mídia do Conhecimento

PTFE - politetrafluoretileno

RNN - *Recurrent Neural Networks*

S&T - *science and technology*

S2ORC - *Semantic Scholar Open Research Corpus*

SA - *Sentiment Analysis*

SBC – Sistemas Baseados em Conhecimento

SBI - *Strategic Business Insights*

SDK - *Software Development Kit*

SESTI - *Scanning for Emerging Science and Technology Issues*

SHRIMP - *Sensitive High-Resolution Ion MicroProbe*

SI - Sistema de Informação

SNA - *Social Network Analysis*

SQUID - *Superconducting Quantum Interference Device*

SVM - *Support Vector Machines*

SVD - *Singular Value Decomposition*

TCP - *Transmission Control Protocol*

TDT - *Topic Detection and Tracking*

TF - *Term Frequency*

TFAMW - Technology Futures Analysis Methods Working Group

TIC - Tecnologias da Informação e Comunicação

TIM - *Tools for Innovation Monitoring*

TOA - *Technology Opportunities Analysis*

TPU – *Tensor Processing Unit*

TRL - *Technology Readiness Level*

ULMFiT - *Universal Language Model Fine-tuning*

USPTO - *United States Patent and Trademark Office*

UPAC - *International Union of Pure and Applied Chemistry*

VoIP - *Voice over Internet Protocol*

VSM - *Vector Space Model*

VUCA - *Volatility; Uncertainty; Complexity; Ambiguity*

WEs - *Word Embeddings*

WoS – *Web of Science*

WSD - *Word Sense Disambiguation*

SUMÁRIO

1 INTRODUÇÃO

1.1 PRÓLOGO

O desejo por antecipação de informações pode ser considerado um traço característico humano. Historicamente, muito se dedicou a observação e identificação de padrões para prever chuvas e estações do ano, por exemplo. De início, esse processo de anunciar "sinais do futuro" era atribuído aos especialistas da época, tal como, pensadores religiosos, padres e xamãs, como também filósofos, escritores de ficção, astrólogos, videntes, dentre outros "profetas não-científicos", que se ocupavam basicamente da observação, imaginação e interpretação dos sinais da natureza (NIINILUOTO, 2001; SON, 2015). Posteriormente, com o progresso, essa aptidão passou a ser utilizada para tentar predizer, a título de exemplo, o mercado de ações e eleições políticas. Sob essa perspectiva, nota-se que a curiosidade para com o futuro é parte da natureza humana e a habilidade de previsão e predição vem sendo moldada às circunstâncias e ferramentas do momento.

Todavia, paralelamente a este desejo, as previsões tendem a ser falhas ou desacreditadas. O mito grego de Cassandra ilustra esse episódio, no qual Cassandra é dotada com a capacidade de prever o futuro, no entanto, é amaldiçoada para que nunca acreditarem nela (JOHNSTON; CAGNIN, 2011). Como a maioria dos mitos, essa passagem tem como propósito elucidar traços da natureza humana. Neste caso, destaca-se como os seres humanos desenvolvem fortes modelos mentais do futuro que são utilizados para tomar decisões (BARNES, 1984). Isso se torna um problema quando em vez de examinar novas evidências e revisar esses modelos mentais, a maioria ou distorce novas informações para corresponder às suas expectativas ou, se as informações não podem ser ajustadas a seu modelo, as ignora completamente (BONACCORSI; APREDA; FANTONI, 2020; ROSSEL, 2012; ROUSSEAU; CAMARA; KOTZINO, 2021).

Contudo, o desejo persiste, em especial, acerca do avanço tecnológico. A preocupação com a evolução das tecnologias e a intenção de predizer sua trajetória data desde a revolução industrial no século XIX, como documenta H. G. Wells em "*Anticipations of the Reactions of Mechanical and Scientific Progress upon Human Life and Thought*" de 1902. Em uma transmissão da rádio BBC (*British Broadcast Corporation*) de 1932, o autor H. G. Wells observou:

> "[...] embora tenhamos milhares e milhares de professores e centenas de milhares de estudantes de história trabalhando nos registros do passado, não há uma única pessoa em qualquer lugar que faça um trabalho permanente de estimar as consequências futuras de novas invenções e novos dispositivos. Não existe um único professor de Antecipação [*foresight*] no mundo" (WELLS, 1987, p. 90, tradução nossa).

Assim, embora, possuir informações com antecedência que permitam planejamento não sejam desejos e necessidades recentes, o campo de *Future Studies* apresenta atual interesse e crescimento (GORDON *et al.*, 2020; MÜHLROTH; GROTTKE, 2018; ROHRBECK; BATTISTELLA; HUIZINGH, 2015; ROUSSEAU; CAMARA; KOTZINO, 2021). Dedicando-se a desenvolver meios para se detectar e interpretar sinais presentes no ambiente artificial[1] vivenciado, prospectando futuros alternativos: prováveis, possíveis e preferíveis.

Muito desse interesse pode-se atribuir a aceleração das mudanças em um ambiente corporativo competitivo, tal que se torna cada vez mais importante o papel da análise das informações para as tomadas de decisão. Nesse sentido, buscaram-se desenvolver ferramentas e técnicas para auxiliar nesse processo de antecipação e planejamento (EEROLA; MILLES, 2011; GIBSON *et al.*, 2018; KAYSER; BLIND, 2017; LEE, 2021; MENDONÇA; CARDOSO; CARAÇA, 2012; ZHAO, TANG, HE, 2023). Face a esse contexto, o avanço da ciência e tecnologia viabilizou o aprimoramento da antecipação de cenários futuros adquirindo estrutura lógica baseada em dados (SON, 2015). Concedida essa breve contextualização do assunto abordado nesta tese, a seguir é aprofundada a temática de sinais fracos como antecipação de informação para o domínio tecnológico.

1.2 CONSIDERAÇÕES INICIAIS

O século XXI apresenta uma conjunção de fatores como: (i) um cenário atual descrito pelo acrônimo VUCA - volatilidade, incerteza, complexidade e ambiguidade (***V**olatility - **U**ncertainty - **C**omplexity - **A**mbiguity)* (BENNETT; LEMOINE, 2014), (ii) um volume de informações disponíveis gerado pelos meios de comunicação eletrônicos (*Big Data*) (CHEN; MAO; LIU, 2014), e (iii) a evolução no processamento computacional ("*CPUs-to-GPUs"* e "2010's *deep learning breakthrough*") (FRADKOV, 2020), que conferem demanda e viabilidade para a operacionalização de análises estratégicas antecipativas baseada em dados

[1] A escolha do termo *artificial* aqui não carrega nenhum tipo de julgamento de valor, como algumas vezes pode ser empregado no sentido de "falso". Mas sim, em contraponto com aquilo que é proveniente da natureza; *natural*. Logo, o artificial surge daquilo que é criado pelo humano.

(GORDON *et al.*, 2020; MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINO, 2021).

Como sociedade ocidental, os conceitos de incerteza e impermanência tipicamente não são internalizados nas reflexões. No entanto, estes se tornam pontos críticos para compreender o século pós-moderno (GORDON *et al.*, 2020; SARDAR; SWEENEY, 2016). Como explicita o conceito VUCA: em tempos de mudança rápida, como o vigente, a incerteza se torna a nova "normalidade" (BENNETT; LEMOINE, 2014). Em decorrência disso, a incerteza deve ser considerada, sobretudo, com o objetivo de não prejudicar a capacidade de planejamento e reação das organizações (BEREZNOY, 2017; ROHRBECK; SCHWARZ, 2013).

Como já apontavam Eisenhardt e Brown, em 1998, uma das maiores ameaças às organizações é não conseguir fazer frente às constantes mudanças que ocorrem no mercado porque não as preveem com antecedência. Mas, como expressam, concisamente Maier *et al.* (2016), em tradução livre, "chutar condições futuras talvez não seja suficiente". Diante dessa realidade, enaltece-se a importância e o interesse por teorias de planejamento estratégico que abarquem o conceito de incerteza e consequentemente, possibilitem evitar ameaças e explorar oportunidades (MENDONÇA; CARDOSO; CARAÇA, 2012; ROHRBECK; BATTISTELLA; HUIZINGH, 2015).

> "Em uma economia onde a única certeza é a incerteza, a única fonte segura de vantagem competitiva duradoura é o conhecimento. (...) Empresas de sucesso são aquelas que consistentemente criam novos conhecimentos, os disseminam amplamente por toda a organização e rapidamente os incorporam em novas tecnologias e produtos" (NONAKA, 1991, p. 96, tradução nossa).

Nonaka (1991), em sua fala icônica, reforça a implicação do conceito de incerteza e acrescenta a importância da busca por conhecimento para a competitividade das organizações. É fato que a informação é um recurso crítico para criar valor, desenvolver e sustentar vantagens competitivas (TEECE; PISANO; SHUEN, 1997). Nesse sentido, na Era do Conhecimento, se faz necessário que as organizações gerenciem as informações do ambiente externo para se manterem competitivas (ALMEIDA; LESCA, 2019; KUMAR; SUBRAMANIAN; STRANDHOLM, 2001; ROHRBECK; SCHWARZ, 2013). Dessa forma, existe um esforço das organizações em aprimorar seus processos de acesso e interpretação de dados, transformando-os em vantagem estratégica (GRIOL-BARR ES; MILLA; MILLET, 2019; SON, 2015).

Todavia, por um lado, o planejamento estratégico está habituado a fazer uso de um tipo de informação conhecido como "sinais fortes", isto é, informações específicas o suficiente,

fatos, para direcionar respostas adequadas (HOLOPAINEN; TOIVONEN, 2012). Contudo, por outro lado, esse cenário dinâmico de rápidas mudanças desperta muitas incógnitas (*'unknows'*), que requerem ferramentas que as traduzam em conhecimento (BURMAOGLU; SARTENAER; PORTER, 2019). Tendo em vista este contexto, uma abordagem estratégica mais flexível, capaz de lidar com a incerteza, considerando fragmentos de informação incipientes, ainda vagos e imprecisos, se torna mais razoável. Nessa linha de raciocínio a teoria de análise de sinais fracos foi proposta (ANSOFF, 1975).

Igor Ansoff (1975), foi um dos primeiros a observar como a miopia inerente das organizações sobre as mudanças no ambiente muitas vezes resultou em oportunidades perdidas e uma falha em responder às ameaças. Essas mudanças, afirmou ele, poderiam ser identificadas e antecipadas por meio da identificação dos chamados 'Sinais Fracos' (termo original, *Weak Signals*), definidos como fragmentos ainda vagos de informação como primeiros indícios de possíveis mudanças futuras. O termo 'Sinal Fraco', atua sucintamente como uma metáfora para lidar com essa necessidade de estar 'antecipado' em relação às mudanças e incertezas do futuro (ROSSEL, 2009).

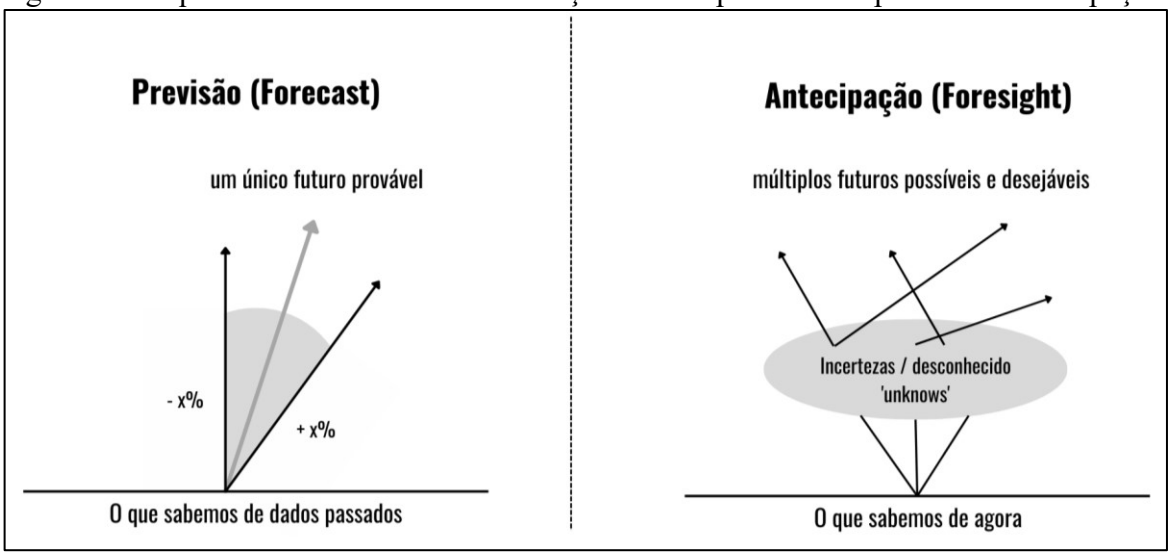
Isto posto, a teoria de sinais fracos, como análise antecipativa, complementa os estudos para planejamento estratégico tradicionais. A abordagem tradicional baseia-se no processo de previsão a partir da projeção de tendência de dados históricos, ou seja, tem como base sinais fortes e pode ser compreendida como parte do campo de estudos de *Forecast*, configurando práticas conhecidas e amplamente empregadas. Por sua vez, a abordagem antecipativa pautada no conceito de sinais fracos, atrela-se à incerteza e se refere à antecipação de informações ainda incipientes, muitas vezes ignoradas pelos modelos de tendência, relacionando-se ao campo de *Foresight*[2] (ALMEIDA; LESCA, 2019). Tem-se então a previsão (*Forecast*) e a antecipação (*Foresight*) como estudos complementares para o planejamento estratégico.

Aprofundando esse entendimento, compreende-se a antecipação como um ato de identificação daquilo que é caracterizado como atual, isso significa que seu ponto de partida é inscrito no presente, e sua intenção voltada para com o futuro e sua construção. Ao passo que as práticas de previsão tradicionais estão relacionadas à projeção de dados passados. Kajikawa *et al.* (2008) trazem essa diferenciação mencionando que enquanto os modelos de antecipação

[2] Adotou-se o termo antecipação como análogo a *Foresight*, ao invés de adotar o termo em português **predição** que pode causar confusão com o termo **previsão,** que por sua vez, está relacionado com a área de pesquisa que *Forecast* que se difere de *Foresight*.

visam identificar os fenômenos em seus estágios iniciais, os modelos de previsão tentam mapear o comportamento futuro dos fenômenos. E, embora o conceito de *Foresight* não seja novidade, apenas recentemente vem se destacando (MÜHLROTH; GROTTKE, 2018; PETER; JARRATT, 2015; ROHRBECK; SCHWARZ, 2013; ZHAO, TANG, HE, 2023), por essa razão, julgou-se necessário estabelecer sua distinção em relação à prática mais conhecida de *Forecast*. A Figura 1 traz um esquema ilustrativo dessa diferença.

Figura 1 – Esquema ilustrativo da diferenciação entre o processo de previsão e antecipação.



Fonte: Elaborado pela autora inspirado em www.futurestation.ro

O embasamento teórico do *Foresight* (COATES; DURANCE; GODET, 2010; MARTIN, 2010; MILES, 2010) reconhece que o futuro não pode ser antecipado de forma linear, sendo crucial estar atento não apenas às tendências (*Forecast*), mas também aos sinais fracos (*weak signals*) e aos cisnes negros (*black swans* ou *wild cards*), isto é, as incertezas ou eventos inesperados que são parte da construção do mundo futuro (HEINONEN *et al.*, 2017). Dessa forma, apenas extrapolar tendências para o futuro não é recomendado devido ao risco de ignorar a não-linearidade, sobreposição ou característica de "surpresa" de alguns fenômenos, ou seja, a necessidade de integrar o aspecto de incerteza no planejamento. (KUOSA, 2010; SARITAS; SMITH, 2011; ROHRBECK; BATTISTELLA; HUIZINGH, 2015).

Nessa perspectiva, como analisa Rossel (2009), Ansoff não foge totalmente da ideia de que o passado é importante para compreender o futuro. Mas, sugere que alguns futuros têm, no momento presente, poucas evidências do que podem vir a se tornar, de modo que são muitas vezes difíceis de detectar. Por conseguinte, as organizações devem aprender a detectar e fazer bom uso dos sinais fracos, antes que o sinal se torne forte (HOLOPAINEN; TOIVONEN, 2012;

ROSSEL, 2009). Dessa maneira, o estudo de previsão de tendências (*Forecasting*) e a detecção de sinais fracos, como matéria-prima para os processos de *Foresight*, se tornam práticas complementares pivotais para um planejamento estratégico robusto (MÜHLROTH; GROTTKE, 2018; PETER; JARRATT, 2015; SARITAS; SMITH, 2011; SCHWARZ, KROEHL; GRACHT, 2014; VEEN; ORTT, 2021).

Diante desse cenário, fica explícito que a detecção antecipada de possíveis mudanças futuras (sinais fracos) é um elemento central para o planejamento estratégico e essencial para que as corporações se mantenham competitivas. No entanto, ainda que a literatura sobre o tema enfatize sobre a importância de sua detecção, e o conceito de sinais fracos tenha sido sugerido há 50 anos, o campo ainda não atingiu maturidade conceitual e estudos reconhecem o estado prematuro das pesquisas e reforçam que há muito para aperfeiçoar tanto nas metodologias, quanto em seus fundamentos teóricos (GRIOL-BARRES; MILLA; MILLET, 2019; LEE; PARK, 2018; MÜHLROTH; GROTTKE, 2018; ROSSEL, 2012; ROUSSEAU; CAMARA; KOTZINOS, 2021; VEEN; ORTT, 2021; ZHAO, TANG, HE, 2023).

1.3 IDENTIFICAÇÃO DO PROBLEMA

Frente a esse contexto, no qual a análise de sinais fracos se mostra cada vez mais pertinente, indaga-se sobre como detectá-los. De modo objetivo, as etapas gerais que descrevem um processo de análise de sinais fracos podem ser compreendidas na: (1) coleta de informações para identificar sinais fracos (*signals*); (2) análise e diagnóstico dos achados (*insights*); (3) formulação da estratégia de ação (*actions*) (VEEN; ORTT, 2021). Ou seja, um encadeamento de operações que visam transformar as informações dos sinais fracos percebidos em conhecimento utilizável.

De início é necessário evidenciar que detectar sinais fracos é uma tarefa intrinsecamente difícil. Por definição, tratam-se de fragmentos vagos de informação que podem facilmente passar despercebidos, normalmente ocultos no ruído dos dados produzidos diariamente. Por ruídos entende-se por informações de fatos irrelevantes que apontam em direções contraditórias (MENDONÇA; CARDOSO; CARAÇA, 2012). Além disso, remetem a eventos sem passado reconhecível, representam uma alteração desconhecida, inesperada ou rara, o que os tornam difíceis de se distinguir como informações relevantes, pois podem parecer aleatórios ou desconexos (ALMEIDA; LESCA, 2019; LEE; PARK, 2018; ROSSEL, 2012; ROUSSEAU; CAMARA; KOTZINOS, 2021; SARITAS; SMITH, 2011; VEEN; ORTT, 2021). Seguindo

esta lógica, justamente por serem difíceis de perceber recebem o nome de ‘fracos’, contudo são relevantes por serem os primeiros indícios de potenciais eventos futuros.

Somado a isso, como mencionado previamente, não apenas as abordagens de identificação necessitam progredir, mas também sua teorização. Apesar de ter seus princípios conceituais bem estabelecidos, sua estrutura teórica do ponto de vista operacional é ainda debatida, sendo às vezes referido como fragmento de informação identificável orientado para o futuro e às vezes como suposições sobre questões emergentes, numa percepção subjetiva (HILTUNEN, 2008; VEEN; ORTT, 2021; ZHAO, TANG, HE, 2023). Essa falta de clareza e consenso em sua definição, como fragmento encontrado ou suposição interpretada, pode ser percebido como um fator agravante para elaboração de meios de identificação desses sinais.

Tradicionalmente, as organizações direcionam especialistas como responsáveis pela coleta e obtenção dessas informações estratégicas. Como consequência, torna-se tênue a distinção entre o processo de identificação (1) e o processo de interpretação dos sinais fracos (2) (ZHAO, TANG, HE, 2023). Sendo a primeira delas, embora de fundamental importância, ainda negligenciada pelas pesquisas (DAY; SCHOEMAKER, 2006; ROHRBECK; SCHWARZ, 2013; ZHAO, TANG, HE, 2023). Nos trabalhos publicados encontram-se muitas abordagens que omitem essa etapa inicial de identificação dos sinais e se ocupam de investigar habilidades e competências das organizações para interpretar, gerir e agir frente a essas informações (ALMEIDA; LESCA, 2019; EEROLA; MILES, 2011; HEINONEN *et al.*, 2017; HOLOPAINEN; TOIVONEN, 2012; PETER; JARRATT, 2015). Nessa linha de raciocínio, compreende-se que as etapas de gestão desse tipo de informação evoluíram ao passo que a fase de coleta e identificação desses sinais desenvolve-se vagarosamente, com poucos modelos se dedicando a isso.

Essa prática mais tradicional, fundamentada em abordagens subjetivas para identificação dos sinais fracos submete as organizações a propensão de negligenciar sinais fracos. É parte do processo cognitivo humano a tendência de atentar-se para aquilo que confirme suas premissas e suposições. Essa ação do viés cognitivo de confirmação pode induzir os especialistas a capturar apenas sinais associados às suas experiências e expectativas anteriores, resultando em métricas falhas e interpretações equivocadas a respeito do futuro (BONACCORSI; APREDA; FANTONI, 2020; EEROLA; MILES, 2011; JOHNSTON; CAGNIN, 2011; GRATCH *et al.*, 2015; KAYSER; BLIND, 2017). Frente a essa realidade, considerando o avanço teórico e computacional para análise de dados, é esperado que decisões

sejam, cada vez mais, tomadas e justificadas baseadas em dados e menos baseadas na intuição de gestores.

Ademais, o crescimento explosivo da disponibilidade de dados impossibilita que esse processo seja executado manualmente, dependendo apenas do especialista. Esse grande volume de dados gerados a cada instante - a era do *Big Data* - revela a necessidade de métodos e técnicas advindas da Computação e Engenharia do Conhecimento (EC) para acessá-los e usar as informações e conhecimento extraídos para apoiar o planejamento e decisões estratégicas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; LIAO; CHU; HSIAO, 2012; SON, 2015; VASSAKIS; PETRAKIS; KOPANAKIS, 2018; ZHAO, TANG, HE, 2023).

Curiosamente, Ansoff em seu trabalho *"Competitive strategy analysis on the personal computer"* de 1986, já reconhecia vantagens em usar ferramentas computacionais para apoiar sua teoria. Na visão ansoffiana, sinais fracos são compreendidos como fragmentos de informação que existem independentes da capacidade de interpretação de um especialista. Assim, apresentam propriedades objetivas e mensuráveis que podem ser detectadas por meio de buscas e análises sistemáticas (ROSSEL, 2012). Sob essa perspectiva, estimula-se que sejam elaboradas estratégias quantitativas de detecção desses sinais, embora ainda este seja um esforço escasso.

Na literatura, embora constem algumas propostas quantitativas, estas essencialmente se fundamentam em abordagens estatísticas de mineração de textos (EBADI; AUGER; GAUTHIER, 2022; GRIOL-BARRES *et al*., 2019; 2020; YOON, 2012; ZHU; DU, 2023). Atualmente, grande parte das informações produzidas estão dispostas de maneira não estruturada, principalmente em texto (CHEN; MAO; LIU, 2014; VASSAKIS; PETRAKIS; KOPANAKIS, 2018). Além disso, considerando que métodos e técnicas de EC mais sofisticadas estão disponíveis e são capazes de explorar semanticamente esses documentos, é esperado o desenvolvimento de abordagens quantitativas que explore capacidades atuais das técnicas de mineração de texto no contexto da detecção de sinais fracos. Particularmente, abordagens que explorem o conteúdo não estruturado de bases textuais, implicando em um viés de análise semântica e extração do conhecimento. Todavia, no presente momento, autores queixam-se da falta de ferramentas para suportar computacionalmente a análise de sinais fracos (AMANTIDOU *et al.*, 2012; GARCIA-NUNES *et al.*, 2020; GORDON *et al.*, 2020; ZHAO, TANG, HE, 2023).

Convém observar ainda, que a fase de coleta é a que mais se beneficia de ferramentas computacionais. Autores apontam que o envolvimento de especialistas seja restrito às fases finais de interpretação/ideação, tomada de decisão e implementação de estratégia, onde o conhecimento humano ainda se mostra valioso e superior ao computacional (AMANTIDOU *et al.*, 2012; MÜHLROTH; GROTTKE, 2018; SARITAS; BURMAOGLU, 2015). O processo de interpretação de sinais fracos e *Foresight* permanecerá uma atividade criativa e centrada no humano, no entanto, cada vez mais as ferramentas computacionais devem servir de apoio (KELLER; GRACHT, 2014; ZHAO, TANG, HE, 2023). Em suas revisões Gordon *et al.* (2020) e Zhao, Tang e He (2023) frizam ainda a importância das técnicas de IA e automação para a encurtar o tempo necessário das organizações em traduzir os sinais em *insights* e ações, avançando para uma tomada de decisões em tempo real. O que reforça a necessidade do desenvolvimento de métodos e técnicas computacionais para a fase de detecção de sinais fracos.

1.3.1 Recorte

Conforme apresentado, os sinais fracos fornecem fragmentos de informação que podem ser usados como gatilhos para decisões estratégicas e, consequentemente, o planejamento antecipado em resposta às ameaças e oportunidades. Esses possíveis eventos/fenômenos futuros podem ser mudanças políticas, novas tecnologias, fenômenos econômicos, novos concorrentes entre outros, comumente referenciados por PEST[3] *check-list* (*politics, economics, society, and technology*) (ANSOFF, 1975; HILTUNEN, 2008; HOLOPAINEN; TOIVONEN, 2012; MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINO, 2021; ZHAO, TANG, HE, 2023).

Dentre estas esferas o domínio tecnológico se destaca. Historicamente, é foco de estudos futuros devido à estreita relação entre progresso tecnológico e o avanço econômico (BUNGE, 1985; GROVER, 2019; ROSENBERG, 1982). Tanto que compreende um nicho de pesquisa dedicado: o *Future-Oriented Technology Analysis* (FTA) (PORTER, 2007; EEROLA; MILES, 2011). Além disso, vale frisar que o momento atual de expansão tecnológica decorrente da Transformação Digital, expresso também como Quarta Revolução Industrial ou Indústria 4.0, enaltece o interesse pela antecipação de informações tecnológicas, uma vez que esse progresso

[3] Ou variações, PESTE (*Political-military, Economic, Socio-demographic, Technological* e *Environmental*), STEEP (*Social, Technological, Economic, Ecological* e *Political*), PESTLE (*Political, Economic, Social, Technological, Legal* e *Environmental*).

contribui para elevação do grau de incerteza no mercado (CHIARELLO *et al*., 2018; GALATI; BIGLIARDI, 2019; MAGRUK, 2021).

Esse contexto de aceleração tecnológica assinala para um ciclo de vida reduzido para muitas tecnologias e para a confluência de descobertas tecnológicas emergentes ocasionadas pela fusão de tecnologias anteriormente separadas, considerando a aproximação do físico e do digital e a crescente interconectividade de ferramentas e máquinas (KIM; LEE, 2017; SCHWAB, 2016). Como destacam Galati e Bigliardi (2019), nesse cenário as tecnologias alavancam umas às outras, amplificando mutuamente seu impacto. Sendo assim, existe uma ampla gama de desafios e oportunidades que estão transformando organizações, processos de negócios e a própria natureza da competição (PIETREWICZ, 2019). Destarte, na era do conhecimento e aceleração tecnológica, a antecipação da informação relaciona-se a uma maior probabilidade de garantir competitividade, mitigar riscos ou aproveitar oportunidades, promovendo destaque para a importância da detecção de sinais fracos no domínio tecnológico.

Em virtude desse cenário de novidades, ressalta-se a antecipação de informações a respeito de tecnologias emergentes como essencial para a gestão estratégica e um diferencial competitivo para as organizações (EULAERTS *et al.*, 2019; LI, 2021; MAGRUK, 2021; RANAEI *et al.*, 2020; YANG *et al.,* 2022). Esse interesse em particular ocorre pelo potencial de impacto das tecnologias emergentes. Como expõem Day e Shoemaker (2000), tecnologias emergentes se diferem de melhorias incrementais em tecnologias estabelecidas por seu potencial para criar novas indústrias ou transformar a indústria existente. Ademais, a importância de uma tecnologia emergente não se deve apenas as novas oportunidades que ela pode oferecer diretamente, mas também a uma infinidade de impactos que podem se originar dela indiretamente (MOEHRLE; CAFEROGLU, 2019). A propósito, em 1902, H. G. Wells já levantava a provocação de que abordar implicações das novas tecnologias de forma sistemática possibilitaria uma sociedade melhor.

Sob esse viés, a literatura exprime o interesse na antecipação de informações sobre novidades tecnológicas e tecnologias emergentes (COZZENS *et al*., 2010; DERNIS; SQUICCIARINI; PINHO, 2016; JOUNG; KIM, 2017; LEE *et al.*, 2018; LI, 2017; MOEHRLE; CAFEROGLU, 2019; MOMENI; ROST, 2016; RANAEI *et al*., 2020; XU *et al*., 2021; ZHOU *et al.*, 2020; ZHU *et al.*, 2019). Contudo, embora o campo de estudos para *Future-Oriented Technology Analysis* seja vasto, e a teoria de sinais fracos faça sentido nesse contexto, não há muitos estudos que se comprometam com este objetivo de maneira explícita.

Do conjunto de referências apenas quatro trabalhos podem ser mencionados estabelecendo uma relação explícita entre sinais fracos e tecnologias emergentes: Amanatidou *et al.* (2012); Eulaerts *et al*. (2019), Yang *et al.* (2022) e Zhan e Du (2023). Contudo, é notável a relevância desses estudos. Os dois primeiros de comissões de pesquisa da União Européia, seguidos pela publicação do centro de pesquisas governamentais da Coreia do Sul e, o mais recente, é uma publicação ainda em *preprint,* evidenciando um interesse que se desponta no tema. Nessa perspectiva, manifesta-se o interesse não apenas em antecipar informações sobre o domínio tecnológico, mas, precisamente, em detectar sinais fracos relativos a novidades tecnológicas, com o intuito de explicitar potenciais tecnologias emergentes.

Ademais, enfatiza-se a relevância da teoria de sinais fracos para o contexto ao salientar que o desenvolvimento tecnológico configura um sistema adaptativo complexo (CAS - *Complex Adaptative System*) (LANSING, 2003; MCCARTHY, 2003). Isso significa que podem ser compreendidos em retrospectiva (*ex-post*) a sobrevivência e adesão de certas tecnologias (‘*survival of the fittest*’), mas, não existe um prognóstico (*ex-ante*) do surgimento da tecnologia mais apta (‘*arrival of the fittest*’). Nesse sentido, devido à incerteza inerente do domínio tecnológico a teoria dos sinais fracos se mostra mais relevante, em comparação a teoria clássica de previsão baseada na projeção de dados passados, para identificar o possível surgimento de uma tecnologia emergente.

Análises de tendências (*Forecasting*) dependem de um conjunto de dados históricos, ou seja, necessitam de uma base de dados suficientemente grande para projetar uma tendência e realizar uma previsão, isso significa que esse tipo de análise contempla desenvolvimentos que já vêm ocorrendo há algum tempo (HILTUNEN, 2008; KUOSA, 2010; SARITAS; SMITH, 2011). São úteis para acompanhar a evolução e monitorar o comportamento de algo que se tem interesse, mas não para investigar o surgimento de algo novo que ainda se desconhece. Sob essa ótica, basear um planejamento estratégico apenas com análises de tendências pode ser considerado desatualizado, tendo em vista a aceleração do desenvolvimento tecnológico. Tal que, esperar até que seja possível prever tendências pode resultar em um atraso significativo para se beneficiar de oportunidades ou se precaver de ameaças advindas das incertezas intrínsecas do desenvolvimento tecnológico.

Dando continuidade, é importante considerar que as descobertas científicas e sua incorporação em tecnologias envolvem processos complexos de pesquisa e desenvolvimento (P&D), podendo durar vários anos (HWANG; SHIN, 2019). Diante do exposto, em vez de

aguardar até que seja possível identificar a tecnologia emergente em si, se mostra promissora a habilidade de investigar sinais fracos para tecnologias emergentes quando se tem interesse em antecipar informações. Nesta linha de raciocínio, a espera pela identificação de uma tecnologia emergente, e não do seu sinal fraco, pode não garantir as vantagens esperadas. Tal que a detecção de sinais fracos supre essa necessidade de identificar um surgimento tecnológico, "o quanto antes" como frisado por Ranaei *et al.* (2020). De modo que os sinais fracos identificados atuam como *proxies*, indícios, sinalizando essas possíveis futuras tecnologias emergentes e, potencialmente, disruptivas, que são valiosas para o mercado.

Em vista desse panorama, evidencia-se a lacuna de modelos para a antecipação de informações acerca de tecnologias emergentes sob o aspecto de sinais fracos, sendo capaz de lidar com grandes volumes de dados não estruturados e múltiplas bases de dados, fornecendo informações "o quanto antes". A concluir, em síntese, são enumerados alguns dos desafios encontrados na literatura para conduzir a detecção de sinais fracos para tecnologias emergentes:

- Volume de dados (ROUSSEAU; CAMARA; KOTZINOS, 2021; ZHAO, TANG, HE, 2023);
- Viés cognitivo (BONACCORSI; APRENDA; FANTONI, 2020);
- Distinguir sinal fraco de ruído (ROUSSEAU; CAMARA; KOTZINOS, 2021);
- Explicitar informações com maior nível de antecedência possível (RANAEI *et al*., 2020);
- Processamento temporal e dinâmico dos dados (MÜHLROTH; GROTTKE, 2018; ZHAO, TANG, HE, 2023);
- Aprofundamento semântico (ROUSSEAU; CAMARA; KOTZINOS, 2021);
- Uso de fontes de informação variadas (MÜHLROTH; GROTTKE, 2018; ZHAO, TANG, HE, 2023).

1.4 PERGUNTA DE PESQUISA

Até este ponto, foram apresentadas algumas das situações que instigam o desejo de previsões por parte do comportamento humano. Objetivamente, este tipo de conhecimento pode ser categorizado em duas vertentes complementares: a que prevê a partir de projeções com base em dados passados; e a que se ocupa da antecipação a partir da detecção de informações do presente, com o intuito de prover indícios sobre possíveis eventos futuros. Esta segunda, se refere aos fragmentos de informações vagos e incipientes, que, no entanto, apresentam

relevância por seu potencial impacto, chamados de sinais fracos. Mais ainda, pontua-se como atualmente, além do fator histórico, o contexto tecnológico demanda a identificação destes sinais fracos, especialmente para tecnologias emergentes, levando ao problema de pesquisa, desta tese:

Como detectar sinais fracos para tecnologias emergentes, a partir de fontes de dados não estruturados, considerando técnicas computacionais para o suporte de processos de planejamento estratégico?

1.5 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos desta tese.

1.5.1 Objetivo Geral

Desenvolver um modelo para detectar sinais fracos para tecnologias emergentes considerando métodos e técnicas de descoberta de conhecimento em texto.

1.5.2 Objetivos Específicos

i. Identificar fatores e padrões envolvidos na detecção de sinais fracos;
ii. Identificar potenciais técnicas de mineração de texto para detecção de sinais fracos;
iii. Estabelecer conceitualmente sinais fracos para tecnologias emergentes;
iv. Demonstrar a capacidade e viabilidade (nível de protótipo) do modelo realizando experimentos em cenários de estudo.

1.6 JUSTIFICATIVA E RELEVÂNCIA DO TEMA

Estudos prospectivos sobre o futuro como o *Foresight* atuam como uma sistematização para avaliar a longo prazo a ciência, tecnologia, economia e sociedade, com o objetivo de determinar as dimensões estratégicas capazes de garantir benefícios socioeconômicos (BEREZNOY, 2017; COATES *et al.*, 2001; SLAUGHTER, 1990). Em outras palavras, o *Foresight* trabalha paralelamente com os conceitos de futuros possíveis e futuros desejáveis. Nessa perspectiva, os resultados da análise de sinais fracos podem ser usados tanto para desenvolver a consciência de ameaças e oportunidades quanto para apoiar a geração de ideias e a criação de conhecimento (ROUSSEAU; CAMARA; KOTZINO, 2021). Dessa forma, o

processo de *Foresight,* e os sinais fracos como seus componentes, criam valor providenciando acesso a recursos críticos antes da concorrência, preparando a organização para mudanças e permitindo que a organização se oriente proativamente em direção ao futuro que almeja (ROHRBECK; BATTISTELLA; HUIZINGH, 2015).

Perante o exposto, esse cenário atual de desenvolvimento acelerado impulsionado pelo avanço das tecnologias, implica em um aumento do aspecto de incerteza nos planejamentos (MAYER *et al.*, 2016; PETER; JARRATT, 2015; RANAEI *et al.*, 2020). Tal que, visualizar posicionamentos futuros apenas a partir da projeção de dados do passado não se mostra suficiente. Assim, destaca-se o interesse no conceito de sinais fracos como teoria de antecipação de informação, visando uma maior a probabilidade de garantir competitividade, mitigar riscos ou aproveitar oportunidades.

Em particular, expressa-se interesse pelo domínio tecnológico, no intuito de aplicar o conceito de sinais fracos para antecipar informações acerca de tecnologias emergentes. Esse interesse se justifica pela relação entre progresso tecnológico e desenvolvimento econômico. Segundo Schumpeter (1943), o crescimento econômico ocorre com a introdução de novos produtos e processos, ou seja, novas tecnologias. Fazendo do campo de estudos tecnológicos uma área fértil e incessante para novas pesquisas, como bem resumem Hussain, Tapinos e Knight (2017, p.160, tradução nossa) "A evolução da tecnologia e a procura pela 'próxima grande coisa' é uma busca constante das organizações". Por consequência, a associação dessa informação ao contexto de desenvolvimento acelerado, resulta no interesse por estudos antecipativos para tecnologia (*Technology Foresight*), em particular pela antecipação de informações acerca de tecnologias emergentes, fornecendo informações essenciais para planejamento e expansão do conhecimento tecnológico (GIBSON *et al.*, 2018).

À luz desses fatos, a obtenção da informação provida pela detecção de sinais fracos para tecnologias emergentes possibilita a ideação de novas capacidades e planejamento visando competitividade. A partir desse ponto de vista, o ambiente altamente competitivo de hoje, pautado pela era do conhecimento, faz da capacidade de uma empresa em buscar informações para acompanhar o progresso tecnológico e inovar continuamente uma habilidade crucial para sua sobrevivência e crescimento (ALMEIDA; LESCA, 2019; KUMAR; SUBRAMANIAN; STRANDHOLM, 2001; ROHRBECK; SCHWARZ, 2013). Compreende-se por inovações como o processo de criação de valor para uma "ferramenta", logo, entende-se que inovações podem surgir a partir de avanços tecnológicos e do surgimento de novas tecnologias, como

novas ferramentas. Assim, o contato com novas informações providas pelos sinais fracos cria espaço para inovação, e inerentemente viabiliza crescimento econômico (MENDONÇA; CARDOSO; CARAÇA, 2012).

Ademais, o interesse pelo domínio tecnológico ocorre por sua capacidade de influência na relação interdependente entre ciência, tecnologia e sociedade. O progresso tecnológico caminha de modo a economizar tempo e insumos fornecendo novas bases de conhecimento acarretando mudanças para os negócios e sociedade (BUNGE, 1985; GROVER, 2019; ROSENBERG, 1982). Novas tecnologias alteram a forma como são executadas nossas atividades, como observa Marshall McLuhan (1962): a tecnologia é criada por nós e então ela nos cria. Por exemplo, a criação da *Internet* definiu a maneira como são realizadas as atividades de trabalho, a busca por entretenimento e até como interações sociais são conduzidas. Nessa conjuntura, ciclicamente nos negócios e na sociedade emergem novas necessidades, demandando mais soluções tecnológicas (BETZ *et al.*, 2019). Semelhantemente, o próprio avanço científico necessita do suporte de tecnologias para avançar. À vista disso, justifica-se a relevância de investigar o domínio tecnológico também pelo seu poder de impacto e influência em demais esferas.

Diante desse cenário, a gestão do conhecimento sobre tecnologias é vital para competitividade tanto ao nível individual de uma organização, quanto nacional para desenvolvimento da nação (EULAERTS *et al.*, 2019; ROHRBECK; SCHWARZ, 2013; SHIBAIYAMA; YIN; MATSUMOTO, 2021). Visto a importância do tema, é pertinente apresentar que grandes programas nacionais e governamentais, como o programa de pesquisa da *US Intelligence Advanced Research Projects Activity* (IARPA) de 2011 - "*Foresight and Understanding from Scientific Exposition* (FUSE)", se comprometem com a mineração de *big data* relacionado à ciência e tecnologia.

Particularmente, no contexto nacional, em sua condição de país em desenvolvimento, Su e Lee (2010), Schoeneck *et al*. (2011) e Pietrobelli e Puppato (2016) mencionam os benefícios de estudos de *Foresight* para o Brasil. Segundo os autores, países em desenvolvimento apresentam recursos limitados e a adequada alocação de recursos se torna crucial, dessa forma estudos de antecipação tecnológica precisam ser conduzidos a fim de sustentar a competitividade nacional em níveis elevados (PIETROBELLI; PUPPATO, 2016; SCHOENECK *et al.*, 2011; SU; LEE, 2010). Tornando-se pertinente disseminar essa área de pesquisa para o desenvolvimento do país.

Por fim, para exemplificar a importância da identificação de sinais fracos para tecnologias emergentes, comenta-se o caso da *American Online*® (AOL). Como premissa, considera-se o fato de que tecnologias se tornam obsoletas, e muitas vezes esse processo ocorre de forma rápida, de modo que, no momento em que uma empresa se compromete com uma grande aplicação de capital em uma tecnologia, de repente, ela pode parar de ser usada, deixando a empresa presa a um investimento sem retorno. No início dos anos 2000, quando a AOL se uniu a *Time Warner* ®, era esperado que elas formassem a maior empresa de mídia do planeta. Mas, a tecnologia de discadores de *internet* que movia a AOL 'subitamente' tornou-se obsoleta. Ou seja, os envolvidos na fusão *AOL-Time Warner* falharam em identificar que uma nova tecnologia surgia - a *internet* banda larga - comprometendo o negócio que ficou conhecido como um dos maiores fracassos corporativos. E, a AOL, que já foi considerada uma das empresas mais tecnológicas não se manteve inovadora e atualmente é pouco conhecida (MALONE; TURNER, 2010).

A concluir, o avanço tecnológico gera oportunidades, riscos e incertezas, a serem exploradas. Sendo assim, um modelo para detecção de sinais fracos para tecnologias emergentes atende o desejo dos estudos de antecipação tecnológica fornecendo informações o mais cedo possível para tomadores de decisão orientarem seus planejamentos estratégicos.

1.6.1 Originalidade da Pesquisa

A originalidade da pesquisa surge da abordagem quantitativa para identificação de sinais fracos com uma proposta baseada em mineração de texto considerando a semântica presente nos documentos. Implementando duas bases como fontes de conhecimento, publicações científicas e pedidos de patentes. Bem como da proposição conceitual de novas palavras como sinais fracos para tecnologias emergentes.

Embora o interesse sobre informações acerca de tecnologias emergentes seja evidente, tendo em vista seu potencial valor para o mercado, os estudos dedicados à coleta de informações tecnológicas se apresentam ainda dependentes da análise de especialistas ou contam com métodos bibliométricos simples, baseados em padrões e frequência, não sendo efetivos para antecipação de informações "o quanto antes".

A contribuição, portanto, a partir da abordagem proposta da descoberta de sinais fracos baseado em novas palavras, oferece fragmentos de informação nova não apenas do ponto de vista de especialistas, mas novas no âmbito acadêmico-científico. Logo, satisfazendo a

demanda pelo maior nível de antecedência possível, colaborando para os estudos de planejamento estratégico sob o domínio tecnológico antecipando informações sobre possíveis tecnologias emergentes.

A complementar, o planejamento estratégico quanto atividade de *Foresight* é a capacidade e habilidade de planejar ações visando o futuro com base no conhecimento adquirido no presente. Perante o fato de que grande parte das informações produzidas estão dispostas de maneira não estruturada, principalmente em texto. Considerando também, o volume de informações, em específico sobre publicações científicas, só tende a crescer. Segundo a pesquisa de Bornmann e Mutz (2015), aproximadamente a cada nove anos, dobra-se o volume de produção científica global. Ademais, a fim de evitar a interferência de vieses cognitivos na identificação dos sinais fracos (BONACCORSI; APRENDA; FANTONI, 2020), o uso de modelos de Descoberta de Conhecimento em Texto (*Knowlege Discovery in Text* - KDT) se mostram adequados para investigar o problema.

Nesse sentido, analisar grandes quantidades de informação para identificar sinais fracos e apoiar estudos de *Foresight* é uma tarefa importante, porém um trabalho moroso. Em razão disso, buscar por meios automáticos ou semiautomáticos para a detecção de sinais fracos se torna relevante. Embora o uso de KDT nesse contexto seja ainda incipiente, a literatura apresenta algumas propostas de modelos para detecção de sinais fracos que se utilizam de modelos de Processamento de Linguagem Natural (NLP), como mencionado anteriormente, os estudos de Eulaerts *et al*. (2019), Yang *et al.* (2022) e outros a serem apresentados na seção de 2.5 de Trabalhos Relacionados. Contudo, existem ainda muitos percalços a serem superados, configurando a lacuna a qual esta tese se propôs a contribuir.

Diante do exposto, evidencia-se a contribuição para a antecipação de informações acerca de tecnologias emergentes sob o aspecto de sinais fracos, sendo capaz de lidar com grandes volumes de dados, múltiplas bases de dados e considerando a semântica contextual das palavras. Visando explicitar as informações com mais antecedência se comparado ao tempo necessário para uma detecção baseada na ocorrência das palavras mais frequentes. Além disso, considerou-se também a questão de temporalidade e dinamicidade por meio de técnicas de análise de redes, aspectos apontados pelos estudos de detecção de sinais fracos como uma deficiência.

1.6.2 Contribuições

- Contribuição acadêmica

Contribui-se para o progresso científico ao passo que propaga-se novos conhecimentos com a produção desta tese, apresentando um novo modelo baseado em uma abordagem ainda não explorada pela literatura. Em especial, destaca-se a contribuição para os estudos de sinais fracos sob o foco da etapa de identificação apontada como deficiente. Apresenta-se um modelo baseado em métodos e técnicas de descoberta em texto, explorando a aplicação de técnicas correspondentes ao estado da arte em NLP, Aprendizado de Máquinas e Análise de Redes, dispondo da capacidade semântica dessas técnicas, identificando palavras novas e estimando seu potencial de impacto para detecção de sinais fracos para tecnologias emergentes.

Existe um grande interesse em informações acerca de tecnologias emergentes tendo em vista seu potencial valor para o mercado. Nesse sentido, muitos estudos demostram interesse por informações antecipadas, mas poucos empregam o conceito de sinais fracos. Dessa forma, colabora-se para os esforços de antecipação de informação tecnológica como aquelas realizadas pela União Europeia com JRC (*Joint Research Centre*) (EULAERTS *et al*, 2019) e pelo Instituto Coreano de Ciência e Tecnologia da Informação - KISTI (YANG *et al*., 2022) explicitando a relação entre os temas de sinais fracos e tecnologia emergente.

Ademais, com o intuito de propor avanços no conhecimento teórico da área, além de metodologicamente, contribui-se refletindo conceitualmente sobre sinais fracos como palavras novas e também sobre a diferenciação entre sinais fracos para tecnologias emergentes e emergência tecnológica.

- Contribuição Gerencial

A estratégia é um plano para atingir um ou mais objetivos em condições de incerteza, os sinais fracos contribuem diretamente para esse cenário, onde existe a dificuldade, e muitas vezes a ausência, de tomada de decisão sob incerteza. Atualmente, dentro da economia global, pautada em informações e dados, organizações de todos os tipos requerem informações estratégicas para garantir que sua tomada de decisão seja competitiva em ambientes complexos e incertos. E, embora seja difícil antecipar avanços tecnológicos, a investigação de sinais fracos para tecnologias emergentes é uma ação importante capaz de oferecer um norte para as decisões organizacionais, delimitando o domínio de informações a serem consideradas por especialistas,

auxiliando no processo de direcionamento sobre onde concentrar recursos e esforços, financeiros e humanos.

Nesse sentido, a detecção de sinais fracos para tecnologias emergentes contribui como ponto inicial para o planejamento estratégico em dois aspectos principais: (1) fornecendo conhecimento empírico necessário para embasar questões de gestão, ao indagar sobre implicações caso o sinal fraco venha a se tornar um sinal forte; e (2) para a criação de novas trajetórias de futuro em um processo de criação de conhecimento, voltado para a ideação de um futuro desejado a partir das informações do presente, promovendo direção e foco para as mudanças, podendo inclusive conduzir ao rompimento de modelos mentais predominantes e encorajar a pensar de forma diferente (GORDON *et al*., 2020; ROHRBECK; SCHWARZ, 2013).

A contribuição gerencial, portanto, ocorre pela antecipação de informações acerca de tecnologias emergentes ao detectar seus sinais fracos concedendo mais tempo para análises e decisões, o que pode resultar em ganhos de competitividade para tais atividades fins. Contribui, particularmente, para pesquisas na área de planejamento estratégico para a ciência e tecnologia (*Science and Technology* - S&T), tendo em vista que o sucesso dos métodos de monitoramento tecnológico (FTA) depende dos dados de entrada, assim uma lista de recomendações de sinais fracos para tecnologias emergentes é um dado de entrada importante para diversas técnicas de *Foresight*.

- Perspectiva de Projeção dos Resultados da Pesquisa para a Sociedade

A suposição fundamental é que o futuro é desconhecido, mas as antecipações e decisões de hoje podem ajudar a preparar e orientar o progresso para o futuro. Junto ao desenvolvimento de novas tecnologias, informações sobre sua viabilidade e implicações associadas permanecem desconhecidas. Considerando a relação e a influência que o desenvolvimento tecnológico tem sob questões socioeconômicas, se torna fundamental a antecedência dessas informações para a geração de políticas e normas futuras sobre os ideais de segurança, transparência, justiça entre outros.

Políticas cientificas definem princípios gerais de operação na ciência em uma sociedade ou nação. Por política entende-se como o conjunto de princípios que definem métodos gerais de ação, responsável pela orientação, incentivo, criação, aquisição, desenvolvimento e disseminação de tecnologia (LUNDVALL; BORRÁS, 2005; MEISSNER; GOKHBERG;

SOKOLOV, 2013). Por conseguinte, oportunizar o acesso à informações antecipativas sobre tecnologias emergentes confere tempo hábil para discussões entre líderes e pesquisadores realizarem um gerenciamento eficaz da tecnologia e formulação de políticas para a Ciência, Tecnologia e Inovação (CT&I). Assim, identificar as melhores práticas necessárias para garantir que a sociedade obtenha os benefícios das tecnologias emergentes enquanto gerencia de forma responsável seus riscos.

A complementar, a antecedência de informações acerca de tecnologias emergentes pode, ainda, auxiliar na transição, repleta de barreiras, da incorporação de novas tecnologias. Sabe-se que pode levar muitos anos até que uma descoberta científica resulte em uma nova tecnologia para o mercado (HWANG; SHIN, 2019). Nesse sentido, acredita-se que, quanto antes se expor novos achados científicos-tecnológicos, incentiva-se o aprofundamento das pesquisas, intencionando diminuir esse intervalo e acelerando o processo de difusão. Assim, dando antecedência para planejamento do meio que essa tecnologia será levada da ciência para o mercado (ERZURUMLU; PACHAMANOVA, 2020). Podendo contribuir até para que, se reconhecido antecipadamente, seja impedido que um achado tecnológico relevante se torne uma '*sleeping beauty*' (RAAN, 2017).

1.6.3 Escopo e delimitações

Como observam Schoemaker e Day (2009), há uma grande diferença entre identificar sinais e perceber o que eles significam. Nessa linha de raciocínio, este estudo se limita ao compromisso de detecção de sinais fracos para tecnologias emergentes. Por conseguinte, etapas posteriores à identificação, como interpretação, análise evolutiva e estratégias de ação fogem do escopo. Friza-se que um sinal fraco é um indício e não uma constatação, existindo um processo de evolução de sinais fracos para fortes, logo este carrega diferentes possibilidades de desenvolvimento; tal que um sinal pode se tornar rapidamente uma tendência; demorar muitos anos para se destacar; pode ainda ocorrer de não gerar grande impacto, ou até mesmo, que seja descartado com o passar do tempo. Assim, o foco na identificação não abarca esses desdobramentos.

Também não se tem a pretensão de tipificar o sinal detectado. A saber, existem estudos que se dedicam a definir e classificar os diferentes tipos de sinais fracos em relação ao seu potencial impacto. Como, por exemplo, diferenciar sinais fracos de *wild cards*. Esta pesquisa, limita-se a desenvolver um modelo focado na etapa de detecção desses sinais fracos, em outras

palavras, na descoberta de conhecimento objetivando fornecer indícios cada vez com mais antecedência referente ao domínio tecnológico.

Outra delimitação ocorre ao se distinguir para esta pesquisa os temas de ‘tecnologias emergentes’ de ‘emergência tecnológica’, muitas vezes tratados como equivalentes nas pesquisas voltadas à FTA. Entende-se, que o termo ‘emergência tecnológica’ esteja mais próximo ao conceito de descoberta de lacunas do conhecimento (encontrar a ausência de informação), tipo de investigação que foge ao escopo deste trabalho focado no termo de ‘sinais fracos para tecnologias emergentes’ referente às informações que já existem, mas encontram-se ainda veladas.

Além disso, uma limitação do estudo decorre em vista de que uma análise completa de *Technology Foresight* deve encontrar um equilíbrio adequado entre ambos fatores “*technolgy/science-push*” e “*demand/market-pull*”. Aqui limita-se a contribuir com informações do tipo “*technology/science-push*” a partir da investigação de publicações cientificas e pedidos de patentes. Compreende-se que tecnologias não surgem apenas a partir de avanços na ciência (ROSENBERG, 1982), mas a ciência é um importante arcabouço para esses indícios, de modo que se justifica investigar descobertas científicas como sinais para tecnologias emergentes. Tem-se o entendimento que o caminho não é exclusivamente esse: ciência, tecnologia, inovação e produtos. Mas este é um dos caminhos possíveis e mesmo limitando-se a ele seu estudo é capaz de contribuir destacando tópicos importantes a serem considerados pelos gestores tomadores de decisão. Ademais, entende-se também, que existe uma estreita relação entre tecnologia e inovação, mas sem o aprofundamento nesta questão limitando-se apenas ao conceito de tecnologia.

Por fim, embora seja uma importante ferramenta, não se aprofunda na questão de visualização da informação (InfoVis). Da mesma forma que não se tem a pretensão de investigar as razões que levaram a existência do sinal fraco para tal tecnologia emergente, e também não se tem o intuito de conjecturar sobre questões éticas ou legais sobre o futuro da tecnologia. Ademais, questões sobre a operacionalização do modelo, sob o ponto de vista computacional, como a mensuração da capacidade de processamento necessária, o planejamento de processamento paralelo de máquinas, entre outras decisões práticas, fogem ao escopo.

1.7 ADERÊNCIA AO PPGEGC

Esta seção cumpre o propósito de contextualizar a tese desenvolvida no Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC). A seção se organiza quanto ao seu objeto de formação e pesquisa (1.7.1 Identidade da Tese), quanto à área de concentração e linha de pesquisa (1.7.2 Contexto Estrutural no EGC) e quanto aos trabalhos pregressos já realizados que se contextualizam nos mesmos fatores de identificação desta tese (1.7.3 Referências Factuais).

1.7.1 Identidade da Tese

O PPGEGC tem como objeto de pesquisa o conhecimento, compreendendo-o de modo interdisciplinar, enquanto conteúdo ou processo resultante de interações sociotécnicas entre agentes humanos e tecnológicos. Sob esse viés, boa parte do conhecimento pode ser considerado presente de modo oculto em dados não estruturados armazenados em algum tipo de mídia (texto, imagens, vídeos, áudio). Nesse sentido, a detecção de sinais fracos como objetivo desta tese se relaciona com o objeto de formação e pesquisa do EGC ao promover a descoberta de conhecimento em texto auxiliando no planejamento estratégico considerando o atual ambiente de incerteza decorrente do acelerado desenvolvimento e avanço tecnológico.

Nessa perspectiva, adota-se uma visão cognitivista sobre o conhecimento, ou seja, entende-se o conhecimento como algo materializável, representável, estocável, replicável e produzível, por agentes humanos ou artificiais (SANTOS; RADOS, 2020). Tal que, a Engenharia do Conhecimento (EC) define metodologias e ferramentas a fim de adquirir e modelar conhecimento e permitir sua apropriação por processos de Gestão do Conhecimento (GC). Por conseguinte, ao desenvolver um modelo de descoberta de conhecimento em texto para detecção de sinais fracos para tecnologias emergentes, o aspecto que contextualiza a tese na área de Engenharia do Conhecimento reside na capacidade do modelo em explicitar conhecimento auxiliando em tarefas intensivas em conhecimento.

Em outras palavras, o resultado proveniente do modelo elaborado como solução, ao apresentar recomendações de sinais fracos para tecnologias emergentes, possibilita a geração de novas ideias, *insights* e novas formas de pensamento. Permitindo que o processo de gestão de conhecimento os utilize em tomadas de decisão para planejamento estratégico, fazendo uso explícito desse conhecimento e gerando valor para organização ou sistema.

1.7.2 Contexto Estrutural no EGC

A Engenharia do Conhecimento se desdobra da Ciência da Computação, como braço da Inteligência Artificial (Artificial Intelligence - AI), e utiliza-se sistemicamente de métodos, técnicas e ferramentas objetivando a extração e representação do conhecimento como suporte para sistemas especialistas (STUDER; BENJAMINS; FENSEL, 1998). Em suma, a EC atua na modelagem do conhecimento e no desenvolvimento de sistemas baseados em conhecimento (SBC) com o intuito de tornar o conhecimento explícito e independente de um intelecto intangível. Facilitando, então, a utilização explícita desse conhecimento, promovendo o compartilhamento e fornecendo insumos para a Gestão do Conhecimento, e assim, gerando valor ao apoiar a resolução de problemas.

Seguindo esse raciocínio, esta pesquisa se contextualiza na área de Engenharia do Conhecimento (EC). Precisamente, na linha de pesquisa Teoria e Prática em Engenharia do Conhecimento, ao fazer uso de ferramental computacional para processos de aquisição e de representação do conhecimento. Dessa maneira, a pesquisa apresentada nessa tese atuou na elaboração, desenvolvimento e implementação de um modelo como proposta de solução de Engenharia do Conhecimento atendendo demandas organizacionais de planejamento estratégico. Como efeito, permite explorar oportunidades de aplicação dos recursos de conhecimento explicitados para processos de engenharia e gestão do conhecimento.

1.7.3 Referências Factuais

- Afinidade por tema:

Os trabalhos do Quadro 1 (Dissertações (D) e Teses (T)) se alinham com o estudo pela temática de inteligência e planejamento estratégico, ou seja, foco na coleta de informações para tomadas de decisão. Cabe observar que nestes casos, os estudos são centrados em inteligência competitiva, ao passo que a tese é voltada para o conceito de inteligência antecipativa.

Quadro 1 – Referências Factuais sobre a temática de Inteligência e Planejamento Estratégico.

Autor	Título	Ano	D/T
Leandro Dal Pizzol	Uso da Web de Dados como Fonte de Informação no Processo de **Inteligência Competitiva** Setorial.	2014	D
Cátia dos Reis Machado	**Análise estratégica** baseada em processos de **Inteligência Competitiva** (IC) e Gestão do Conhecimento (GC): proposta de um modelo.	2010	T
Renata J. Vieira	Incorporação da **Inteligência Competitiva** às Atividades de **Planejamento Estratégico** do Projeto de Produtos Industriais.	2009	T
Paulo Henrique de Souza Bermejo	**Planejamento Estratégico** de Tecnologia da Informação com Ênfase em Conhecimento.	2009	T

Fonte: Elaborado pela autora

Por sua vez, os trabalhos do Quadro 2 se alinham com a tese sob a ótica de interesse no domínio tecnológico. Apresentam pesquisas direcionadas à tomada de decisão para ciência e tecnologia e abordam o contexto do avanço tecnológico.

Quadro 2 – Referências Factuais sobre a temática de Tecnologia.

Autor	Título	Ano	D/T
Paulo Cesar Lapolli	Estratégias para a concepção de competências essenciais à luz do sistemismo no contexto da **indústria 4.0**	2022	T
Leandro Quingerski	KE-IOT: Uma proposta de modelo de sistema baseado em conhecimento para ambientes de **Internet das Coisas** (IoT)	2019	D
Divino Ignácio Ribeiro Jr.	Modelo de sistema baseado em conhecimento para apoiar processos de **tomada de decisão em ciência e tecnologia**	2011	T

Fonte: Elaborado pela autora

- Afinidade por problema de pesquisa (aplicação):

As pesquisas apresentadas no Quadro 3 têm como foco a descoberta de conhecimento em texto. Em especial alinham-se com esta tese os estudos de Woszezenk (2016) e Bovo (2011), por trazerem o conceito temporal para a análise.

Quadro 3 – Referências Factuais sobre o problema de pesquisa: Descoberta de Conhecimento em Texto.

Autor	Título	Ano	D/T
Marina Carradore Sérgio	**Modelo** de Avaliação de Potenciais **Ideias** Alinhadas ao Contexto Organizacional.	2020	T
Alessandro Costa Ribeiro	Modelo de reconhecimento de padões em ideia usando técnicas de **descoberta de conhecimento em textos.**	2018	D
Cristiane Raquel Woszezenk	Modelo Para **Descoberta de Conhecimento** Baseado em **Associação Semântica e Temporal** Entre **Elementos Textuais.**	2016	T
Dhiogo Cardoso da Silva	Uma arquitetura de **business intelligence** para processamento analítico baseado em **tecnologias semânticas e em linguagem natural**.	2011	S
Alessandro Botelho Bovo	Um modelo de **descoberta de conhecimento** inerente à **evolução temporal** dos relacionamentos entre **elementos textuais.**	2011	T

Fonte: Elaborado pela autora

- Afinidade por método/abordagem adotado:

Quanto a processo metodológico adotado, a *Design Science Research Methodology* (DSRM) é a principal abordagem das dissertações e teses do programa em Engenharia do Conhecimento, por seu embasamento para o desenvolvimento de artefatos. Consoante ao propósito de pesquisas tecnológicas (a metodologia DSRM é apresentada em detalhes na seção 3.2).

Por fim, a complementar, enaltece-se a relevância e atualidade da temática proposta por esta tese em sintonia com a publicação sobre sinais fracos por dois doutorandos e professores do programa, Machado et al. (2020). Essa tese, portanto, se apresenta alinhada e contributiva para o histórico do Programa, ocupando-se de um tema latente ainda pouco explorado pelo PPGEGC.

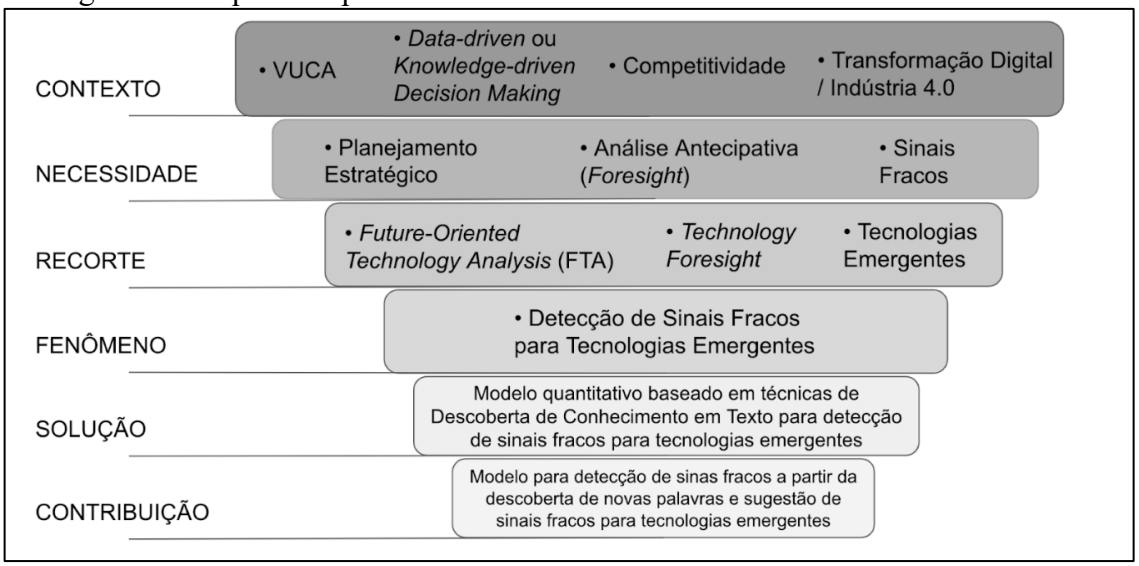
1.8 ESTRUTURA DO TRABALHO

Esta tese segue um fluxo encadeando os conhecimentos como mostra o esquema da Figura 2. O documento da tese inicia com a Introdução no Capítulo 1, apresentando o contexto de mudança, resumidos pelo acrônimo VUCA, e marcado por questões como a Economia do Conhecimento, enaltecendo as informações como um bem de valor e consequentemente a análise de dados como um direcionador na tomada de decisões. Esse cenário destacou a

importância e necessidade do planejamento estratégico pelas organizações, em especial, modelos que abarquem a incerteza como a abordagem de sinais fracos, elemento do processo de *Foresight*.

Fazem parte da motivação também, o contexto de Transformação Digital (ou Indústria. 4.0) ressaltando ainda mais o já existente interesse nas pesquisas voltadas ao domínio tecnológico e nas tecnologias emergentes. Contudo, a detecção dos sinais fracos é ainda apontada como uma dificuldade e abordagens computacionais quantitativas de dados não estruturados seguem pouco exploradas para antecipação de informação tecnológica, principalmente sob o viés de exploração da semântica presente nos textos, sendo esta, portanto, a lacuna a qual se dedica esse trabalho.

Figura 2 – Esquema representando o fluxo de conhecimentos envolvidos na tese.



Fonte: Elaborado pela autora

O segundo capítulo, refere-se à Fundamentação Teórica da tese apresentando a literatura mobilizada a fim de resolver o objetivo proposto. Constando o histórico e definições dos principais assuntos da tese: sinais fracos; estudos antecipativos para tecnologia, descoberta de conhecimento em texto e a investigação do processo de surgimento de terminologias científicas. Em seguida, o Capítulo 3 dedica-se aos procedimentos metodológicos da pesquisa incluindo a caracterização do estudo e descrição da *Design Science Research Methodology* escolhida para embasar a elaboração do modelo proposto.

No Capítulo 4 encontram-se os resultados da tese. Contendo o embasamento teórico das premissas do modelo proposto, a apresentação do modelo e o racional de seu desenvolvimento, seguido de uma instanciação demostrativa para ilustrar o funciomento do modelo. A fim de

evidenciar as proposições conceituais de sinais fracos como palavras novas e o entendimento sobre sinais fracos para tecnologias emergentes como resultados da pesquisa teórica elas se localizam nesse capítulo de resultados junto a apresentação do modelo. O Capítulo 5, tem por finalidade avaliar o modelo proposto de acordo com sua efetividade como solução para atender o problema identificado, assim traz a operacionalização do modelo simulando sua aplicação.

Considerações finais, ponderando limitações e passos futuros, constam no Capítulo 6. A listagem de Referências encerra este documento. O Quadro 4 sintetiza essas informações.

Quadro 4 – Síntese do conteúdo dos capítulos da tese.

Capítulo	Descrição do Conteúdo
1. Introdução	Este capítulo apresenta um panorama da pesquisa. Inicialmente aborda o contexto em termos do interesse e importância da antecipação da informação, em especial, dos sinais fracos, como fragmentos ainda vagos de informação. A introdução segue ressaltando a relevância das pesquisas antecipativas para o domínio tecnológico. Na sequência, são apresentados o objetivo geral e os objetivos específicos. Em seguida a justificativa com argumentos relativos ao seu ineditismo e à sua contribuição. Por fim, é apresentada a organização deste documento.
2. Referencial Teórico	Os conceitos apresentados neste capítulo fundamentam a pesquisa envolvendo o histórico e definição de sinais fracos, estudos antecipativos para tecnologia e descoberta de conhecimento em texto. Investigando também a formação de novas palavras como aspecto central para a proposta de solução do modelo de detecção de sinais fracos para tecnologias emergentes. O capítulo finaliza com os principais trabalhos relacionados ao problema de pesquisa declarado.
3. Método de Pesquisa	O capítulo se dedica a elucidação sobre a visão de mundo a partir da qual a pesquisa se fundamenta e sua classificação como pesquisa tecnológica. Apresenta a metodologia *Design Science Research Methodology* (DSRM) utilizada e descreve os procedimentos metodológicos realizados.
4. Modelo Proposto	Esse capítulo apresenta o resultado principal da tese, o artefato desenvolvido, no caso, um modelo para detecção de sinais fracos para tecnologias emergentes. Contendo também a conceituação teórica que embasa o modelo e a demostração do modelo passo-a-passo em um cenário fictício
5. Avaliação Experimental	Nesse capítulo a análise do modelo ocorre a partir da simulação operacional em um caso em retrospectiva.
6. Considerações Finais	Capítulo que encerra o documento da tese sintetizando principal contribuição e direcionando próximos passos da pesquisa.

Fonte: Elaborado pela autora

2 REFERENCIAL TEÓRICO

Este capítulo abarca a literatura visitada que alicerça a pesquisa, passando pelos seguintes assuntos: seção 2.1, apresenta o histórico do conceito de sinais fracos e suas definições; seção 2.2, explora historicamente os estudos antecipativos para o domínio tecnológico; seção 2.3 compreende o campo de estudo de Descoberta de Conhecimento em Texto (KDT), explorando abordagens específicas para representação vetorial das palavras e análise de redes complexas temporais; seção 2.4 investiga como ocorrem as nomenclaturas científicas, tendo em vista a proximidade entre ciência e tecnologia e o interesse na descoberta de novas tecnologias. Por fim, a seção 2.5 compila os trabalhos relacionados os quais fornecem importantes contribuições para o modelo desenvolvido e a seção 2.6 recapitula o encadeamento lógico da fundamentação teórica da tese.

2.1 SINAIS FRACOS

O conceito de sinais fracos surge proposto por Ansoff (1975) ao complementar a importância do monitoramento do ambiente (*environmental scanning*), ou monitoramento da informação, apresentado por Aguilar (1967). Igor Ansoff é tido como pai do planejamento estratégico, foi um matemático aplicado russo-americano por formação, mas com interesses interdisciplinares. Em seu trabalho no departamento de matemática da *RAND Corporation*® (empresa que oferece pesquisas e análises para as Forças Armadas dos Estados Unidos) foram iniciados os estudos sobre a sua percepção a respeito da miopia organizacional das instituições face às descontinuidades, o que se tornou o foco central de seus estudos. Em sua publicação *"Managing Strategic Surprise by Response to Weak Signals"* (1975), Ansoff argumenta sobre a implementação de um sistema organizacional de alerta precoce às mudanças, os sinais fracos ou do termo original em inglês, '*weak signals*'.

Estudos estratégicos tiveram um desenvolvimento acentuado no contexto militar, após a Segunda Grande Guerra e principalmente no período da Guerra Fria. Posteriormente, esse conhecimento foi adaptado para o contexto organizacional de planejamento estratégico corporativo (ROSSEL, 2012). Ansoff traduziu em uma proposta metodológica a ideia da necessidade de estar 'antecipado', ou melhor, 'saber o mais cedo possível', em relação a uma mudança, fornecendo à esfera corporativa uma abordagem capaz de lidar com as constantes incertezas e mudanças do mercado e do mundo. A visão estratégica é importante considerando a vantagem competitiva concedida ao propiciar a possibilidade de aproveitar oportunidades e

mitigar riscos antecipadamente (MENDONÇA; CARDOSO; CARAÇA, 2012; VEEN; ORTT, 2021).

A escolha da expressão sinal fraco por Ansoff, foi inspirada na tradição cibernética, na teoria da informação de Shannon (1948) e em analogia a ideia dos radares (ROSSEL, 2012). Em sua proposição sinais fracos atuam como avisos de potenciais eventos futuros (externos ou internos a uma organização) que ainda são muito vagos para permitir uma estimativa precisa de seu impacto e/ou capaz de prover uma resposta absoluta de reação. Precisamente, descrito conceitualmente por Ansoff e McDonnell como:

> "um acontecimento sobre o qual apenas informações parciais estão disponíveis no momento em que uma resposta precisa ser tomada, para que seja concluída antes que o acontecimento tenha impacto na empresa" (ANSOFF; MCDONNELL, 1990, p.490, tradução nossa).

Nessa perspectiva, um sinal fraco apresenta três principais características na informação que carrega: novidade, incerteza e relevância estratégica (ANSOFF, 1975; ROSSEL, 2009, 2011, 2012; VEEN; ORTT, 2021). Logo, sinais fracos são considerados 'fracos' não por falta de importância, mas, porque são facilmente ofuscados por outros fatores. São fragmentos incompletos de difícil identificação e interpretação podendo passar facilmente despercebidos, embora potencialmente importantes. Nas palavras de Mendonça, Cardoso e Caraça (2012), em uma produção abundante de estímulos, em meio a ruídos, pode haver informações incipientes, incompletas, não estruturadas e fragmentadas, apontando para o surgimento de transformações desafiadoras.

Os sinais fracos contrapõem os sinais fortes (*strong signals*), aquelas informações sobre as quais se tem convicção. Estabelecendo um paradoxo estratégico entre esperar ter informações suficientes para tomar decisões ou aceitar uma informação ainda vaga e agir se ajustando à imprecisão inerente a ela, tal que não seja tarde demais para tomada de decisões estratégicas (MENDONÇA; CARDOSO; CARAÇA, 2012; PETER; JARRATT, 2015). Neste sentido Ansoff sugere o monitoramento contínuo a partir da identificação de um sinal fraco, para que os tomadores de decisão (gestão estratégica) consigam planejar uma resposta gradual e evolutiva conforme o desenrolar dos sinais fracos.

Em virtude da atual aceleração nos desenvolvimentos e mudanças devido ao progresso tecnológico (BENNET; LEMOINE, 2014; MAGRUK, 2021), evidenciou-se o aspecto da incerteza nos planejamentos estratégicos. Tal que, Ilmola e Kussi (2006) descrevem esse momento como o "renascimento" da análise de sinais fracos, tendo em vista que o conceito foi

proposto há mais de quatro décadas. Os autores, acentuam quão atraente tem sido o legado de Ansoff e também quão diversificada se tornou a perspectiva do conceito de sinal fraco.

A análise de sinais fracos é uma abordagem que pode detectar indícios de mudanças futuras em estágio inicial. Deste modo, sinais fracos podem remeter a um cenário geral a partir da análise conjunta de diferentes esferas do PESTEL (*Political, Economics, Societal, Technological, Environmental and Legal*) (HILTUNEN, 2008; MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINO, 2021). Como também, remeter a cada uma dessas esferas individualmente, compreendendo fragmentos de informações sobre uma mudança política, uma nova tecnologia, um fenômeno econômico, um novo competidor, entre outros. Contemplando ainda, visões micro, meso ou macro, ou seja, para uma organização em particular, um setor/nicho específico ou ao nível de nação (ANSOFF, 1975; HILTUNEN, 2008; HOLOPAINEN; TOIVONEN, 2012; MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINOS, 2021).

A relação entre sinais fracos e o domínio tecnológico se torna evidente sob a perspectiva que sinais fracos representam informações de relevância estratégica, com potencial, porém, de impacto desconhecido, assim o interesse pelas informações tecnológicas se justificam, tendo em vista que avanços tecnológicos repercutem na economia e sociedade (BUNGE, 1985; ROSENBERG, 1982; SCHUMPETER, 1943). Nesse sentido, relatórios governamentais evidenciam o interesse e a relevância estratégica das análises de sinais fracos para o domínio tecnológico, como o projeto do CGEE[4] - Centro de Gestão e Estudos Estratégicos do Brasil, com a publicação "Metodologia para identificação de sinais fracos e monitoramento de tendências globais em CT&I" (CGEE, 2016; MORESI, 2019), os estudos sobre sinais fracos coordenados pelo KISTI[5] - *Korea Institute of Science and Technology Information* (KWON *et al.*, 2018; LEE *et al.*, 2018; YOO; WON, 2018; YANG *et al.*, 2022), o projeto da União Europeia SESTI[6] (*Scanning for Emerging Science and Technology Issues*) (AMANATIDOU *et al.*, 2011), e a pesquisa '*Weak Signals in Sciente and Technologies*' (EULAERTS *et al.*, 2019; 2021; 2022) da comissão europeia JRC[7] - *Joint Research Centre of the European Commission.*

[4] CGEE - https://www.cgee.org.br/
[5] KISTI - https://eng.kist.re.kr/kist_eng/main/
[6] SESTI - https://cordis.europa.eu/project/id/225369/reporting
[7] JRC - https://publications.jrc.ec.europa.eu/repository/handle/JRC119395

Convém observar, que embora a compreensão do conceito de sinais fracos e sua importância seja clara, a definição apresentada por Ansoff (1975) é ampla e não constitui uma descrição operacional. Além disso, as possibilidades de aplicação do conceito estabelecem uma nomenclatura diversa (HOLOPAINEN; TOIVONEN, 2012; VEEN; ORTT, 2021). Por conseguinte, apesar de Ansoff ter iniciado a conceitualização de sinais fracos para o planejamento estratégico, vários outros acadêmicos buscaram aprimorar sua teoria e diferentes tentativas de definir esse conceito vem sendo discutida ao longo do tempo (HILTUNEN, 2008; KUOSA, 2010; MÜHLROTH; GROTTKE, 2018; ROSSEL, 2012; ROUSSEAU; CAMARA; KOTZINOS, 2021; SARITAS; SMITH, 2011). Como efeito, constitui-se uma lista de termos relacionados como: *weak signal* (sinal fraco), *early signal* (sinal inicial), *future sign* (signo futuro), *warning signal* (sinal de alerta), *early warning signal* (sinal de alerta inicial), *novel future signal* (sinal novo futuro), *disruptive signals* (sinais disruptivos), *drivers of change* (condutores de mudanças), *seeds of change* (sementes de mudança), entre outros. Em meio a esses termos, evidencia-se a contribuição de Hiltunen (2008), ao presentar a estruturação teórica de '*future sign*', com ampla adoção nos trabalhos de sinal fraco, sistematizado a partir das dimensões da semiótica de Peirce.

Diante do exposto, é possível encontrar na literatura diversas definições para terminologia de sinal fraco, como se pode ser visto em Veen e Ortt (2021) que revisitaram 68 definições de sinais fracos para a proposta de uma definição unificada. Todavia, uma característica unânime entre os pesquisadores sobre sinais fracos é o fato de ser o primeiro indício sobre uma possível mudança futura, sendo os sinais fragmentos de informação aparentemente insignificantes, mas que podem fornecer *insights* sobre possíveis eventos futuros. Como desfecho, a proposta de definição unificada apresentada por Veen e Ortt (2021, p.11, tradução nossa) segue: "Sinal fraco é uma percepção de fenômenos estratégicos detectados no ambiente ou criados durante a interpretação que estão distantes do quadro de referência do observador"[8].

Em sua proposta os autores evidenciam que novidade, relevância estratégica e incerteza são condições necessárias, embora não sejam suficientes para estabelecer um sinal fraco. No sentido que, se um sinal que não é novo (inédito/desconhecido), mas tem importância estratégica e é percebido como uma ameaça ou oportunidade para um observador, passa a ser

[8] Trecho original: "*A weak signal is a perception of strategic phenomena detected in the environment or created during interpretation that are distant to the perceiver's frame of reference*." (VEEN; ORTT, 2021, p.11).

considerado um sinal fraco sob sua perspectiva. Essa ponderação vai ao encontro a reflexão levantada por Kim e Lee (2017), em seu conceito de '*novel future signal*', que embora todo sinal fraco trate de novidade, alguns sinais fracos estão longe de serem novos (inéditos), contudo, sinais novos são sempre sinais fracos pelo aspeto inerente de novidade.

Com objetivo de aprofundar a discussão sobre a definição unificada apresentada por Veen e Ortt (2021) explicitam-se os estudos de Rossel (2012) e Ahlqvist e Uotila (2020) que discorrem sobre os fundamentos teóricos implícitos na compreensão do conceito de sinais fracos. Para Rossel (2012) existem duas abordagens principais: aquela que encara os sinais fracos objetivamente e outra que os considera subjetivamente. A visão objetiva compreende que o sinal fraco independe de uma interpretação para existir, é um fenômeno em si. Sendo possível identificá-lo sem saber exatamente o que está sendo procurando. Como exprime a definição de Veen e Ortt "Sinal fraco é uma percepção de fenômenos estratégicos detectados no ambiente [...]". Nessa linha de raciocínio, pode-se ou não conseguir identificá-los, mas o fato de não ser capaz de interpretá-los não significa que não existam. Rossel (2012) denomina essa visão objetiva de Ansoffiana.

A segunda, Rossel (2012) denomina de Construtivista, a qual considera a ambiguidade e incompletude intrínseca do sinal fraco, implicando na dependência da capacidade de interpretação do observador para a existência do sinal, o que é essencialmente subjetivo. De modo que sinais fracos não são identificados, mas construídos. Como também é representado na definição de Veen e Ortt "Sinal fraco é uma percepção de fenômenos estratégicos [...] criados durante a interpretação [...]". Agregam ainda para a discussão Ahlqvist e Uotila (2020) apresentando uma visão do conceito de sinal fraco sob uma teoria relacional. A qual destaca a natureza do contexto e do referencial de conhecimento na identificação de sinais fracos, onde para os autores, os sinais fracos nem sempre dizem respeito a questões emergentes (fenômenos futuros), mas podem ser o resultado de uma mudança contextual, uma expansão da perspectiva ou uma mudança na posição do observador.

Por conseguinte, compreende-se que o entendimento de sinais fracos oscila segundo o grau de novidade envolvido (inédito ou novidade sob nova perspectiva), o grau de incompletude (quão vago é o sinal, relativo ao quão fácil ou difícil se torna sua interpretação) e sobre sua origem (identificado ou construído). Nesse sentido, ao passo que, para alguns autores sinais fracos estão associados a descoberta de informações, o sinal como uma ocorrência, fragmentos de informação a serem descobertos em meio a um grande volume de conteúdo produzido, cuja

interpretação é passível de gerar um aviso. Por outro lado, outros autores, consideram sinais fracos quando existe um processo conjunto de interpretação a partir da percepção de um observador. Ressalta-se que não é negada a importância da etapa de interpretação para o processo de análise de sinais fracos, mas nota-se a distinção das definições como um fator importante para a tarefa de detecção dos sinais fracos.

Sobretudo, independente da abordagem, é importante frisar que a análise de sinais fracos não tem a pretensão de fornecer uma "resposta certa" sobre o que fazer ou ser uma previsão profética. Mas, de ser um estímulo para o pensamento, atuando para criar perspectivas e possibilidades que delineiam como o futuro e o ambiente competitivo podem progredir. Nesse sentido, a detecção de sinais fracos atuam como componentes nos processos de *Foresight* (MENDONÇA; CARDOSO e CARAÇA, 2012), o qual compreende que existem múltiplos futuros possíveis, e precisamente qual desses futuros será atingido depende em parte das decisões tomadas agora (GORDON *et al.*, 2020; IDEN; METHLIE; CHRISTENSEN, 2017).

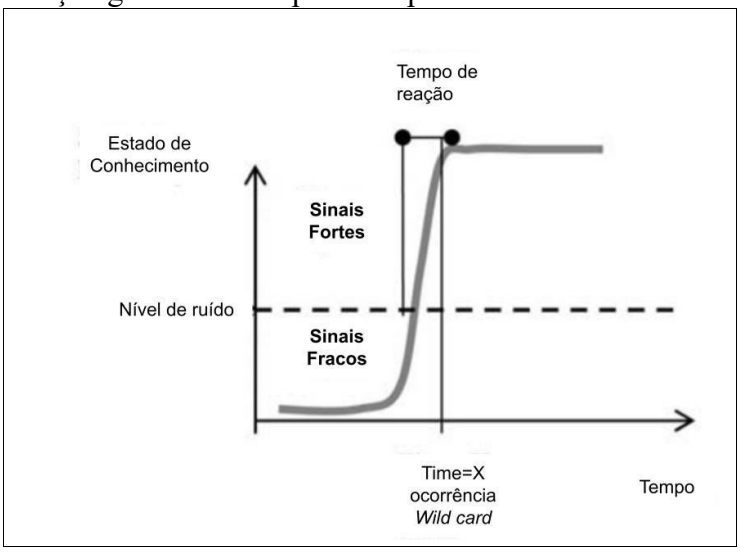
No contexto apresentado, compreende-se que os sinais fracos não são conclusivos, mas são informações de natureza antecipatória que desempenham um papel desencadeador, induzindo o estímulo de indagação a partir de sua detecção ou interpretação e análise sob determinada perspectiva. Somado a isso, cabe ressaltar que se torna demasiadamente complexo o conceito de prever para um sistema que reage a previsões, o CAS (HARRIS; ZEISLER, 2002). Ou seja, diferentemente de uma previsão meteorológica, na qual o palpite não afeta a ocorrência, o fato de antecipar informações como os sinais fracos influencia nas ocorrências subsequentes. São sistemas evolucionários abertos que estão continuamente processando e incorporando ou adaptando-se as novas informações (AALTONEN; SANDERS, 2006).

Nesse sentido, embora não seja possível prever o futuro, é possível desenvolver antecipação e estar ciente daquilo que está surgindo antes que oportunidades sejam perdidas e riscos ocorram. Portanto, a análise antecipativa por meio de sinais fracos não trata como se o futuro estivesse enviando sinais de um futuro preciso. Mas sim, envolve uma atitude conscientemente ativa em relação ao futuro, explorando sistematicamente esses futuros alternativos, reconhecendo que as escolhas feitas hoje podem moldar ou mesmo criar o futuro. Dito isso, o enfoque recai menos sobre a precisão da previsão, e mais na importância de adquirir informações e habilidades para tomar decisões mais inteligentes no presente.

Convém observar ainda, que Ansoff reforça a compreensão da separação entre o que é o sinal de um fenômeno do fenômeno em si (HOLOPAINEN; TOIVONEN, 2012). Visto que um sinal fraco pode indicar diferentes possíveis eventos futuros, como trazer informações sobre tendências futuras, apontar mudanças ou o surgimento de algo novo como fenômeno emergente, bem como, um sinal fraco pode ser um indício de um evento futuro que não venha a ocorrer. Dessa forma, o sinal fraco não é o evento futuro, mas um indício deste, de maneira que os sinais fracos assumem características distintas e particulares em comparação ao fenômeno ao qual se referem (fumaça e fogo). Todavia, mesmo que a diferença entre um sinal e seu fenômeno seja clara na teoria, muitas vezes é difícil realizar essa distinção na prática para sua identificação.

Ademais, julga-se relevante diferenciar a compreensão sobre sinais fracos e o conceito de *wild cards*. O termo *wild cards* foi estudado por Petersen (1999) e definido como eventos com baixa probabilidade e alto impacto, que, caso ocorram, impactariam severamente a condição humana. O conceito é semelhante à teoria do Cisne Negro descrita por Taleb (2007). O que diferencia um sinal fraco e uma *wild card,* como explica Hiltunen (2006), reside no fato de afetarem o futuro de maneira diferente no que diz respeito aos aspectos do tempo e do estado de conhecimento.

Figura 3 – Representação gráfica do tempo de resposta de um sinal fraco e de uma *wild card.*



Fonte: Adaptado de Hiltunen (2006)

Como mostra a Figura 3, um sinal fraco se desenvolve de maneira gradual enquanto uma *wild card* tem um "efeito surpresa", o indício e a ocorrência são imediatos. Nesse sentido, estratégias diferentes devem ser tomadas para se detectar *wild cards* de sinais fracos, pois apresentam características estruturais diferentes. Não se assume que modelos para detecção de sinais fracos não sejam capazes de detectar *wild cards*, e vice-versa, mas não há o compromisso

de um modelo de detecção de sinais fracos em identificar *wild cards*. No entanto, reforça-se que ignorar um sinal fraco pode levar a ter que lidar com um *wild card* e seus impactos (MOHAMMADI; EIVAZI; SAJJADI, 2017).

Prosseguindo, um processo de análise de sinais fracos para antecipação estratégica pode ser compreendido em três principais etapas: a identificação dos sinais fracos (*signals*); o acompanhamento e interpretação desses sinais (*insight*) e; o plano de ação a ser tomado (*action*) (VEEN; ORTT, 2021). Tradicionalmente, metodologias qualitativas como a técnica de cenários (SCHOEMAKER, 1995), a criação coletiva de sentido (LESCA, 1995; LESCA; LESCA, 2011), com a abordagem *LESCAnning*® de Inteligência Estratégica Antecipativa coletiva (IEAc) (JANISSECK-MUNIZ; LESCA; FREITAS, 2011) e o método Delphi (ROWE; WRIGHT, 1999) são usados tanto para a fase de coleta dos dados quanto para as análises e interpretações. Hiltunen (2008) mostra, que até então, humanos eram a fonte de preferência para obter sinais fracos.

Como consequência, o processo para identificação de sinais fracos dependia muito da visão intuitiva de especialistas experientes e da elicitação desse conhecimento por meios qualitativos. Nesse contexto, a obtenção de dados para fomentar a análise antecipativa se torna cara, lenta, pouco disponível e subjetiva. Além disso, com o passar dos anos e o aumento exponencial de disponibilidade de informação, tornou inviável delegar e confiar apenas em especialistas para identificar sinais fracos (MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINOS, 2021; ZHAO, TANG, HE, 2023). Diante desse cenário, métodos baseados em dados possibilitam superar algumas das limitações sofridas pelo processo baseado em especialistas.

Primeiro, tendo em vista que grande parte das informações produzidas estão dispostas de maneira não estruturada, principalmente em texto. Métodos e técnicas de mineração de texto, processamento de linguagem natural e aprendizado de máquina permitem automação na recuperação, extração e análise de documentos, contribuindo para a agilidade do processo e minimizam a questão de subjetividade inerente ao processo qualitativo de identificação dos sinais (LEE; PARK, 2018; MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINOS, 2021). Não descarta-se o uso de especialistas no processo, mas uma combinação de esforços, onde orienta-se que esforços humanos sejam restritos a fase de interpretação e que métodos quantitativos, sejam aplicados para a obtenção das informações para posterior análise, amplificando a capacidade de atuação dos especialistas.

No entanto, nota-se na literatura uma escassez de abordagens operacionalmente embasadas, evidenciando a dificuldade principalmente na fase de detecção dos sinais fracos por sua característica de informação vaga e/ou velada e que se confunde com os ruídos. O fato é que os ruídos são facilmente distinguidos dos sinais fracos em retrospectiva, mas são difíceis de dissociá-los em processos de antecipação (MENDONÇA; CARDOSO; CARAÇA, 2012). Assim, embora haja um consenso sobre a importância estratégica destas informações antecipadas, há dificuldades inerentes à identificação desses sinais.

Sob essa perspectiva, investigaram-se abordagens quantitativas propostas para identificação de sinais fracos. Para esse levantamento foi essencial o trabalho de Rousseau, Camara e Kotzinos (2021) no qual apresentam uma taxonomia sobre todos os métodos utilizados para obtenção de sinais fracos para diferentes domínios. A partir desse estudo atualizou-se a compilação desse conjunto de abordagens quantitativas para identificação de sinais fracos, com ênfase no domínio tecnológico comtemplando o interesse da tese. Nesse sentido, destacam-se as abordagens estatísticas, baseadas na frequência da palavra-chave (Quadro 5), os métodos de monitoramento de tópicos (*Latent Semantic Indexing - LSI e Latent Dirichlet allocation - LDA*) e clusterização (Quadro 6) e abordagens diversas (Quadro 7), como identificação de *outliers*[9].

Sobre as abordagens estatísticas (Quadro 5) destaca-se a abordagem sugerida por Yoon (2012) adotada como base por diversos outros estudos. A proposta fundamenta-se no modelo tridimensional de signos futuros de Hiltunen (2008) e assume como premissa que um sinal fraco é um termo com baixas frequências de ocorrência de documentos e uma alta taxa de crescimento. O autor explora técnicas de mineração de texto e propõe duas métricas: Grau de Visibilidade (DoV - *degree of visibility*), que corresponde à frequência de ocorrência de uma palavra-chave em um conjunto de documentos, e o Grau de Difusão (DoD - *degree of diffusion*), que representa a frequência de documentos de cada palavra-chave. Com essas medidas são gerados mapas de portfólio de palavras-chave: KEM (*Keyword emergence map*) e KIM (*Keyword issue map*), com o propósito de ajudar os especialistas a localizar tópicos de sinal fraco a partir das palavras-chave definidas e distingui-los de demais ruídos.

[9] Um dado observado que destoa significativamente do padrão formado pelas outras observações.

Quadro 5 – Coletânea de estudos baseados na frequência de palavras para a detecção de sinais fracos via mineração de texto relacionados ao domínio tecnológico.

Autores	Ano	Título	
	Abordagem		Base de dados
	Aplicação		
Yoon	2012	*"Detecting weak signals for long-term business opportunities using text mining of Web news"*	
	Grau de Visibilidade (DoV) e Grau de Difusão (DoD)		Web news
	Sinais fracos para ajudar especialistas a reconhecer potenciais oportunidades no ambiente dos negócios		
Kim; Park e Lee	2016	*" A visual scanning of potential disruptive signals for technology roadmapping: investigating keyword cluster, intensity, and relationship in futuristic data"*	
	Frequência de palavras-chave; *self-organising feature map*		*Technology foresight websites*
	Sinais fracos sobre o futuro como apoio ao *roadmapping* tecnológico		
Park e Cho	2017	*" Future sign detection in smart grids through text mining"*	
	DoV; DoD		Radian6
	Sinais fracos para *smart grids* (redes elétricas)		
Lee e Park	2018	*"Simulation of Weak Signals of Nanotechnology Innovation in Complex System"*	
	DoV; DoD; modelo de tópicos		Artigos científicos - EBSCO *database*
	Sinais fracos para questões éticas envolvendo inteligência artificial		
Yoo e Won	2018	*"Simulation of Weak Signals of Nanotechnology Innovation in Complex System"*	
	Infometrics, séries temporais de co-ocorrência de palavras		Artigos científicos
	Sinais fracos de mudança na ciência e tecnologia		
Griol-Barres; Milla e Millet	2019	*"Improving strategic decision making by the detection of weak signals in heterogeneous documents by text mining techniques"*	
	DoV; DoD; análise multi-palavra; NLP		Artigos científicos, notícias de jornal e mídia social
	Sinais fracos para o futuro		
Roh e Choi	2010	*"Exploring Signals for a Nuclear Future Using Social Big Data"*	
	DoV; DoD; análise multi-palavra; NLP		Web news
	Sinais fracos para gestão de informação aplicados à energia nuclear coreana		
Griol-Barres *et al.*	2020	*"Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing"*	
	DoV; DoD; análise multi-palavra; NLP		Artigos científicos, notícias de jornal e mídias sociais
	Sinais fracos para o futuro ajudando especialistas em seus processos de tomada de decisão		
Ebadi; Auger e Gauthier	2022	*"Emerging Technologies and their Evolution using Deep Learning and Weak Signal Analysis "*	
	DoV; DoD; NLP (BERT)		Artigos científicos
	Sinais fracos para tecnologias emergentes no campo da hipersônica		
Zhan e Du	2023	*"Early Identification Methods for Emerging Technologies Based on Weak Signals"*	
	DoV; DoD; LDA; NLP (BERT)		Artigos científicos e patentes
	Sinais fracos para tecnologia emergente de realidade aumentada		

Fonte: Elaborado pela autora

Para os estudos baseados em modelo de tópicos (Quadro 6), destacam-se os trabalhos de Dirk Thorleuchter com a proposta de considerar a semântica dos sinais com foco na análise evolutiva dos sinais fracos. Os autores usam *Latent Semantic Indexing* (LSI) com base em

Singular Value Decomposition (SVD) para criar *clusters* (agrupamentos) analisando as relações semânticas entre documentos. LSI é uma técnica que captura a estrutura semântica latente de uma coleção de documentos, mesmo que não compartilhem exatamente as mesmas palavras ou frases. E SVD é a técnica matemática implementada para redução de dimensionalidade permitindo a operacionalização da abordagem LSI.

Quadro 6 – Coletânea de estudos baseados em modelo de tópicos para a detecção de sinais fracos via mineração de texto relacionados ao domínio tecnológico.

Autores	Ano	Título	
	Abordagem		Base de dados
	Aplicação		
Thorleuchter e Poel	**2013**	***"Weak Signal Identification with Semantic Web Mining"***	
	Modelo de tópicos - *Latent Semantic Indexing* (LSI) e *Singular value decomposition* (SVD)		Internet
	Identifica sinais fracos da internet para uma determinada hipótese, ajudando planejadores estratégicos a reagir com antecedência.		
Thorleuchter; Scheja e Poel	**2014**	***"Semantic Weak Signal Tracing"***	
	Series temporais; LSI; SVD		Internet
	Rastrear o desenvolvimento de sinais fracos ao longo do tempo		
Thorleuchter e Poel	**2015**	***"Idea mining for web-based weak signal detection"***	
	LSI		Internet (*websites*, *blogs*, etc.)
	Sinais fracos como suporte para decisões estratégicas		
Adil e Abdelhadi	**2021**	***"A Framework for Weak Signal Detection in Competitive Intelligence using Semantic Clustering Algorithms"***	
	LDA *(Latent Dirichlet allocation)*; LSA (*Latent Semantic Analysis*) e k-*means cluster*		Artigos científicos e patentes
	Detecção de sinal fraco em grandes volumes de dados para suporte estratégico		
El Akrouchi; Benbrahim e Kassou	**2021**	***"End-to-end LDA-based automatic weak signal detection in web news"***	
	Modelo de tópico - (LDA)		*Web news*
	Sinais fracos para antecipar riscos e oportunidades para o mercado		

Fonte: Elaborado pela autora

No Quadro 7, agrupou-se abordagens distintas. Com menos foco nas técnicas, ressaltam-se as contribuições teóricas da abordagem metodológica do estudo NEST - *New and Emerging Signals of Trends* (Kim *et al*. 2013) por sua robustez e o trabalho de Kim e Lee (2017) pela discussão conceitual sobre antecipação (*earliness*) e novidade (*novelty*) acerca dos sinais fracos.

Quadro 7 – Coletânea de estudos de detecção de sinais fracos via mineração de texto relacionados ao domínio tecnológico.

Autores	Ano	Título	
	Abordagem	Base de dados	
	Aplicação		
Kim *et al.*	**2013**	***"NEST: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network"***	
	Bayesian network	*Global Trend Briefing* (GTB)- KISTI *database*	
	Sinais fracos de tendências emergentes para *Foresight* tecnológico		
Kim e Lee	**2017**	***"Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals"***	
	LOF - *Local Outlier Factor*	*technology foresight websites*; base de patentes	
	Sinais fracos de novidade para tecnologia		
Griol-Barres *et al.*	**2021**	***"Variational Quantum Circuits for Machine Learning. An Application for the Detection of Weak Signals"***	
	Aprendizado de máquina em computação quântica	artigos científicos, notícias de jornal e postagens de mídia social	
	Sinais fracos para o futuro		
Miao; Guo e Yuan	**2021**	***"Research on Identification of Potential Directions of Artificial Intelligence Industry From the Perspective of Weak Signal"***	
	Outliers; modelo de regressão; análise de redes	BVD-*zephyr database*	
	Sinais fracos para a indústria de inteligência artificial se posicionar estrategicamente		

Fonte: Elaborado pela autora

Como conclusão, face às abordagens analisadas para identificação de sinais fracos, observa-se sob uma perspectiva crítica, o desafio de identificar novidade e antecipação em processos computacionais muitas vezes baseados em identificação de padrões. A respeito das principais práticas implementadas, abordagens estatísticas não são capazes de incluir a semântica contida nos textos, se limitando a proporção de palavras-chave, uma abordagem superficial do conteúdo não estruturado presente nos textos. Analogamente, a modelagem de tópicos também se concentra em um padrão, e se caracteriza pela abstração de um conjunto de palavras que são semanticamente próximas umas das outras em um único 'rótulo'. Logo, essas abordagens inviabilizam a identificação da novidade como inédita, do novo 'o quanto antes'.

No momento presente, técnicas de Mineração de Texto e NLP evoluíram significativamente ofertando abordagens mais sofisticadas para se explorar os conteúdos semânticos. Nessa perspectiva, pontua-se que embora os estudos de Ebadi, Auger e Gauthier (2022) e Zhan e Du (2023), por exemplo, façam uso de técnicas de *Deep Learning* para representação do texto, implementando o modelo BERT, o fazem apenas para representar palavras-chave, e seguem com uma abordagem estatística (DoV; DoD) para sugerir sinais fracos. Deixando de explorar a possibilidade de extrair conhecimento das bases não estruturadas

de dados como a semântica contextualizada dos textos. Na seção 2.6 de trabalhos relacionados, estende-se a discussão sobre os principais estudos que contribuem para o objetivo proposto.

2.2 ESTUDOS ANTECIPATIVOS PARA TECNOLOGIA

A Revolução Industrial ofereceu ao mundo moderno novos artefatos e tecnologias. Como consequência, estudos sobre o avanço tecnológico se tornaram um ponto de interesse (LINSTONE, 2011; LINSTONE; TUROFF, 2011). Historicamente, a relevância sobre informações tecnológicas tem sua jornada associada às pesquisas militares para planejamento estratégico, principalmente os estudos desenvolvidos em *Technology Foresight* pelos Estados Unidos da América (EUA) com a *RAND Corporation*, criada em 1948 para oferecer pesquisas e análises às Forças Armadas dos EUA. Na época, no entanto, o termo mais empregado para esse tipo de pesquisa era *Technology Forecasting* (LINSTONE, 2010; MARTIN, 2010; MILES, 2010; ROSSEL, 2012).

Paralelamente, na França, cresciam também os estudos antecipativos para planejamento estratégico sob o nome de *La Prospective* (COATES; DURANCE; GODET, 2010; GODET, 2000). Contudo, enquanto os estudos dos EUA focavam em questões tecnológicas para suas análises futuras (vislumbrando como *seria* o futuro), a França apresentava uma vertente social em suas análises (vislumbrando como *poderia ser* o futuro) (GODET, 2000; ROHRBECK; BATTISTELLA; HUIZINGH, 2015). Com a evolução dos estudos, essas duas vertentes convergiram sob a denominação de *Foresight*, referindo-se às pesquisas antecipativas, englobando ambas reflexões de descoberta e criação de conhecimento, explorando futuros possíveis e desejáveis. O campo de estudo progrediu e se especializou, tal que quando voltado para tecnologia denomina-se *Technology Foresight,* quando adaptado ao contexto organizacional denomina-se *Corporate Foresight* ou *Strategic Foresight* entre outras nomenclaturas que podem ser encontradas na literatura.

Uma definição usual adotada para antecipação tecnológica (*Technology Foresight*) é concedida por Martin (1995):

> "[...] o processo envolvido na tentativa sistemática de olhar para o futuro a longo prazo da ciência, tecnologia, economia e sociedade com o objetivo de identificar as áreas de pesquisa estratégicas e tecnologias genéricas emergentes susceptíveis de produzir os maiores benefícios econômicos e sociais." (MARTIN, 1995, p.140, tradução nossa).

Como um marco importante na literatura de antecipação tecnológica menciona-se o estudo realizado em 1970, pela agência de ciência e tecnologia do Japão (*National Institute of Science and Technology Policy* - NISTEP[10]) para previsão de tecnologia em um horizonte de 30 anos (BREINER; CUHLS; GRUPP, 1994), dando início a um ciclo de estudos de antecipação em nível nacional que atualmente encontra-se em sua 11ª edição. Em 1972, os EUA criaram o escritório *Office of Technology Assessment* (OTA) responsável por fornecer análises oficiais sobre a complexidade do avanço científico e tecnológico do século XX. Desde então, órgãos públicos e privados reconhecem a importância destas análises sob o viés da ciência e tecnologia, passando a incorporá-las em seus planejamentos estratégicos.

Atualmente, os EUA contam com o escritório de estudos de inteligência IARPA (*Intelligence Advanced Research Projects Activity*) e o projeto FUSE[11] (*Foresight and Understanding from Scientific Exposition*), além das pesquisas da *Rand Corporation*[12] e do *Strategic Business Insights* (SBI), com o processo *Scan*®[13]. Esses esforços para monitoramento e antecipação de questões tecnológicas estratégicas são práticas reproduzidas também por alguns governos europeus - Alemanha (*Bundesministerium für Bildung und Forschung* - BMBF[14]) (CUHLS *et al.*, 2009), Reino Unido (*Horizon Scanning Programme Team* - HSPT[15]) (GEORGHIOU, 1996), União Europeia, SESTI (AMANATIDOU *et al.*, 2011) e JRC (EULAERTS *et al.*, 2019; 2021; 2022), entre outros (MARTIN, 2010; MILES, 2010; SON, 2015).

Esse interesse das nações mais desenvolvidas em estarem na vanguarda da ciência e da tecnologia impulsiona a concepção de diversas abordagens para coleta e análise do máximo de informações possíveis para nutrir seus planejamentos estratégicos. À vista disso, o campo de estudo FTA (*Future-Oriented Technology Analysis*) tem por objetivo reunir as diferentes práticas e atividades a fim de analisar o desenvolvimento da tecnologia e suas consequências (HAEGEMAN *et al.*, 2013). Assim, um dos principais focos do FTA é projetar para um

[10] NISTEP - https://www.nistep.go.jp/en/?page_id=56
[11] IARPA - https://www.iarpa.gov/research-programs/fuse
[12] RAND - https://www.rand.org/randeurope/research/futures-and-foresight-studies.html
[13] SBI - http://www.strategicbusinessinsights.com/scan/process.shtml
[14] BMBF - https://www.bmbf.de/bmbf/de/home/home_node.html
[15] HSPT - https://www.gov.uk/government/groups/futures-and-foresight

horizonte futuro, tendo por base o conhecimento disponível sobre questões acerca da ciência e tecnologia, orientando as tomadas de decisão (CAGNIN; HAVAS; SARITAS, 2013; TFAMW GROUP, 2004). Nesse sentido, a atividade do FTA abrange tanto os estudos de projeção de dados passados (*Forecast*) e prospecção a partir de dados presentes (*Foresight*) (EEROLA; MILES, 2011; HAEGEMAN *et al.*, 2013; PORTER, 2007).

Destacam-se, como alguns dos objetivos específicos das práticas FTA: acompanhar o desenvolvimento de certa tecnologia (*technology monitoring*, *technology watch*); coletar e fornecer informações sobre tecnologia (*technology intelligence*, *competitive intelligence*); prever o comportamento futuro da tecnologia (*technology forecasting*); traçar futuros alternativos para o desenvolvimento tecnológico (*environmental/horizon scanning*); avaliar e explorar as tecnologias potenciais que uma organização deve adotar (*technology assessment, technology roadmapping*) ou buscar por indícios iniciais de novidades tecnológicas (*technology alerts/warnings*, *technology foresight*), entre outros (DAIM; OLIVER, 2008; PORTER, 2010; HAEGEMAN *et al.*, 2013). Sob essa análise, pode-se compreender que os estudos para FTA tem interesse em informações sobre três principais pilares: (i) para monitoramento e ponderação da tecnologia, (ii) para incorporar as informações em desenvolvimentos futuros e (iii) para antecipação tecnológica.

Considerando esses propósitos, a literatura para planejamento estratégico atendendo o domínio tecnológico é repleta de estudos voltados à trajetória de evolução das tecnologias, recorrendo a conceitos como *Gartner Hype Cicle*® e *Technology Readiness Level*® (TRL), por exemplo (BILDOSOLA *et al.*, 2017; EFIMENKO; KHOROSHEVSKY, 2017; ENA *et al.*, 2016; LIU *et al.*, 2020; WANG; LIU; LIU, 2017; ZHANG *et al.*, 2016a). Contudo, esses estudos não abarcam a incerteza que um sinal fraco para tecnologia abrange, ainda que tratem de estágios iniciais de desenvolvimentos tecnológicos, tratam-se do estudo de tecnologias já existentes. Por sua vez, os sinais fracos carregam um grau de incerteza antecedente ao TRL1[16], onde princípios básicos da tecnologia são conhecidos, pois, os sinais fracos são indícios que apontam para possíveis tecnologias, sendo que tal tecnologia pode nem mesmo vir a se confirmar. Além disso, as técnicas são focadas em investigar a tecnologia, recordando, que o sinal fraco de um possível evento futuro e o evento futuro em si apresentam características distintas, necessitando técnicas particulares a cada um para sua devida investigação.

[16] https://www.nasa.gov/general/technology-readiness-level/

Em todo caso, esse interesse em investigar a trajetória de evolução das tecnologias é marcado na literatura pelo evidente interesse na antecipação de informações tecnológicas, em particular a respeito de tecnologias emergentes. Sendo recorrente o emprego de expressões como "*early detection*", "*early identification*", "*earliness*" e "*antecipation*", que traduzem esse desejo (HWANG; SHIN, 2019; KUUSI; MEYER, 2002; LEE *et al*., 2018; PORTER; CHIAVETTA; NEWMAN, 2020; WANG, Z. *et al.*, 2019; ZHANG *et al.*, 2014; ZHOU, Y. *et al*. 2019). Ou ainda, expressam essa demanda por antecipação de informações tecnológicas mesmo ao empregarem nos estudos a nomenclatura de previsão e tendência (*Forecasting*) (BILDOSOLA; GONZALEZ; MORAL, 2017; DAIM *et al.*, 2006; KIM; SOHN, 2020; LEE; AHN; KIM, 2021; XU *et al.*, 2021; YOON; LEE, 2012; ZHANG *et al.*, 2016b; ZHOU *et al.*, 2020). São vastos também, estudos que se comprometem a investigar exclusivamente a base de patentes com o intuito de prever ou identificar informações promissoras sobre novas tecnologias (JOUNG; KIM, 2017; KIM, BAE, 2017; KIM; PARK; YOON, 2016; PARK; YOON, 2018; SARICA; LUO; WOOD, 2020; SONG; KIM, Karp Soo; LEE, 2017; SONG; KIM, Kyuwoong; LEE, 2018; MOEHRLE; CAFEROGLU, 2019; ZHOU, X. *et al.*, 2019; ZHOU *et al.*, 2020; ZHU *et al.*, 2019).

No entanto, apesar de manifestarem o interesse na antecipação da informação, grande parte desses estudos se dedicam, principalmente, as atividades de coleta de sinais fortes ou tendências (PETER; JARRATT, 2015). Para a coleta das informações utilizam-se de técnicas qualitativas, como a análise Delphi, ou de técnicas quantitativas, como aquelas baseadas na mineração de texto em patentes e publicações científicas (SARITAS; BURMAOGLU, 2015).

Na evolução de estudos quantitativos voltados a coleta de informações tecnológicas se destacam os estudos de *Topic Detection and Tracking* (TDT) oferecendo métodos sistemáticos para descobrir tópicos em um fluxo textual de notícias (ALLAN *et al.*, 1998), fomentando os estudos de Kontostathis *et al.* (2004) relacionado a Detecção de Tendências Emergentes (ETD) e os inúmeros estudos coordenados por Porter e sua equipe (BURMAOGLU *et al*., 2019; CARLEY *et al*., 2018; COATES *et al.*, 2001; COZZENS *et al*., 2010; GOMILA *et al*., 2021; GUO *et al*., 2012; HUANG *et al.*, 2014; HUANG *et al.*, 2015; LAHOTI *et al*., 2018; LI; PORTER; SUOMINEN, 2018; MA *et al.*, 2014; NEWMAN *et al.*, 2014; PORTER, 2010; PORTER *et al*., 1991; PORTER *et al*., 2019; PORTER; ROSSINI, 1977; RANAEI *et al*., 2020; SCHOENECK *et al.*, 2011; WANG *et al*., 2015; WANG *et al.*, 2017; WATTS; PORTER, 2001; WATTS; PORTER, 2007; YAU *et al*., 2014; ZHANG *et al.*, 2014; ZHANG *et al*., 2016a,b; ZHU; PORTER, 2002).

Com ênfase para a proposição da abordagem *Tech-Mining*, constituída da combinação de técnicas de bibliometria, análise de patentes e diferentes ferramentas de processamento de linguagem natural (NLP) para coletar, processar e visualizar dados sobre tecnologias e os sistemas de gerenciamento de tecnologia como o *Technology Opportunities Analysis* (TOA) (PORTER; CUNNINGHAM, 2004; PORTER; DETAMEPEL, 1995; PORTER; ZHANG, 2015).

Aprofundando a compreensão sobre a obtenção de informações tecnológicas, pontua-se que para cada uma das finalidades do FTA é preciso coletar dados que fomentem a formação de conhecimento. Essa fase de coleta de dados é conhecida como Inteligência. Pode-se dizer, que apesar dos avanços computacionais no esforço da coleta de informações que esses estudos supracitados apresentam, a coleta se concentra em sinais fortes sobre tecnologia, ação conhecida por *Technology Intelligence*. Por outro lado, quando se busca por sinais fracos para tecnologia, o objetivo principal não é identificar e acompanhar os desenvolvimentos tecnológicos ao longo do tempo, mas sim encontrar a "agulha no palheiro", um fragmento de informação nova em grandes coleções de documentos para apoiar a processos de *Foresight*, tarefa que pode ser descrita como *Anticipatory Intelligence*. Refletindo sobre a distinção no tipo de informação que se está interessado, torna-se necessário explicitar a distinção desse processo em *Technology Intelligence* e *Anticipatory Intelligence.*

Para elucidar essa distinção, reforça-se que ao passo que a inteligência tecnológica (*Technology Intelligence)*, busca reunir o máximo de informações possíveis existentes a respeito de determinada tecnologia (PORTER; ZHANG, 2015). No entanto, algumas informações são difíceis de serem identificadas, como os sinais fracos, por configurarem uma forma diferente de informação, sendo necessárias abordagens distintas para sua identificação e análise. Por conseguinte, essa etapa focada na coleta de fragmentos de informação ainda vagos seria adequadamente denominada de inteligência antecipatória (*Anticipatory Intelligence)*, embora não seja uma nomenclatura muito adotada pelos estudos de planejamento estratégico (COATS, 2019). Contudo, pode-se notar o esforço na atualização das abordagens visando analisar tecnologias futuras e suas consequências, evoluindo e amadurecendo ao longo do tempo, mas ainda são poucos os esforços explícitos de identificação de sinais fracos para tecnologia.

A importância dos sinais fracos para a antecipação tecnológica decorre do fato dos sinais fracos se definirem por fragmentos ainda vagos de informação e existir uma noção de progresso

evolutivo do sinal, tal que podem tanto apontar para uma tecnologia que desponte como uma tendência tecnológica futura, quanto podem nem mesmo vir a se concretizar como um sinal forte. Considerando esse conceito e tendo em vista o cenário acelerado de desenvolvimento, a antecipação da informação se mostra valiosa para o planejamento, possibilitando que se desfrute de oportunidades e se previna riscos. Quando se tem o interesse em estar antecipado em relação às informações tecnológicas, mais antecipado do que buscar sobre informações iniciais acerca de uma tecnologia está em buscar sinais fracos sobre tecnologias. Particularmente, no cenário tecnológico interessa-se por sinais fracos para tecnologias emergentes, uma vez que podem apontar para uma tecnologia disruptiva, o que é de grande valor para o mercado (DOTSIKA; WATKINS, 2017; MOMENI; ROST, 2016).

Ademais, o cenário de acelerado desenvolvimento tecnológico atual, pontuado pela quarta revolução industrial (indústria 4.0), assinala para a eclosão de tecnologias emergentes (GALATI; BIGLIARDI, 2019; MAGRUK, 2021). De modo que a investigação sob a perspectiva dos sinais fracos se torna relevante, contribuindo para a conscientização sobre ameaças e oportunidades tecnológicas ou na geração de ideias para planejamento estratégico de tecnologia. Auxiliando na tomada de decisões sobre investimentos em P&D, apoiando pesquisadores e organizações o mais cedo possível na decisão de investir ou abandonar novos desenvolvimentos (AMANATIDOU *et al.*, 2012; EULAERTS *et al*., 2019; YANG *et al.*, 2022).

Convém salientar que estudos que buscam antecipar informações tecnológicas sob o viés de identificar tecnologias emergentes em si se diferem de identificar um sinal fraco para uma tecnologia emergente, pois um sinal fraco não tem o compromisso de atestar todas as características de uma tecnologia emergente (COZZENS *et al.*, 2010; ROTOLO; HICKS; MARTIN, 2015; SUOMINEN; NEWMAN, 2017). Sob essa óptica, embora, muitos apontem a importância dos estudos de antecipação de informações para tecnologia, do ponto de vista de se coletar e incluir os sinais fracos nos processos, ainda persiste a necessidade de desenvolvimento dos estudos voltados à coleta de sinais fracos para tecnologia.

As pesquisas em bases teóricas e técnicas para antecipar o futuro com base em evidências quantitativas de *big data* ainda estão em estágio inicial (RANAEI *et al*., 2020). O estudo recente desenvolvido pelo *Joint Research Centre of the European Commission* (JRC) intitulado "*Weak Signals in Science and Technology*" (EULAERTS *et al*, 2019) é uma das primeiras grandes contribuições alinhada com esse propósito. Repercutindo inclusive como

influência para atualização da abordagem do KISTI (*Korea Institute of Science and Technology Information)* (YANG *et al.*, 2022). Nos quais de forma explicita relacionam os termos sinais fracos e tecnologias emergentes.

Dito isso, ao passo que o planejamento estratégico objetiva antecipar ao máximo as informações, a linha entre pesquisas para antecipação da tecnologia e para fronteira científica se torna tênue. Tanto que o estudo JCR (EULAERTS *et al*, 2019) se coloca como 'detecção de sinais fracos de tecnologias emergentes ou novos tópicos científicos'. Afinal, a descoberta de tecnologias emergentes depende significativamente de avanços científicos (MARTIN, 1995). Assim como define Bunge (1985), a tecnologia é caracterizada pela produção de algo artificial, isto é, de um artefato, envolvendo necessariamente embasamento científico. Sob esse viés, o desenvolvimento científico pode ser considerado como 'sementes' de tecnologia e inovação no modelo de inovação linear (SHIBATA; KAJIKAWA; SAKATA, 2010). Diante dessa perspectiva, considerando o comprometimento em buscar meios de antecipar informações, julgou-se válido explorar também estudos que se comprometem em investigar a fronteira da ciência.

As publicações científicas atuam como importantes portadoras de informações sobre tecnologia sendo um importante recurso de dados para estudar o desenvolvimento e a mudança da tecnologia. Por sua vez, as patentes fornecem uma fonte de informação atualizada e confiável para refletir o desenvolvimento tecnológico. Frente ao desejo de antecipar informação tecnológica, existe uma classe de estudos que se dedicam a investigar a evolução dos desenvolvimentos científicos ou tecnológicos, buscando estabelecer a fronteira do conhecimento, explorando inclusive conjuntamente essas bases de ciência e tecnologia (ÁVILA-ROBINSON; MIYAZAKI, 2013; BA; LIANG, 2021; I'ANSON, 2017; FLEMING; SORENSON, 2004; LI *et al*., 2019; LI *et al*., 2020; RANAEI *et al.*, 2020; SHEN; WANG; YANG, 2020; SHIBATA; KAJIKAWA; SAKATA, 2010; SMALL; BOYACK; KLAVANS, 2014; UPHAM; SMALL, 2010; XU *et al.*, 2019; XU, H. *et al.*, 2020; WANG, X. *et al*., 2019; WU *et al*., 2010; ZHOU Y., 2019).

Todavia, ainda que essas pesquisas reconheçam a importância dessas informações para o planejamento estratégico, não contribuem diretamente para o problema de sinais fracos (CHEN, 2006; DU; WU, 2018; PONOMAREV *et al.*, 2014; SAVOV; JATOWT; NIELEK, 2020; SHIBAIYAMA; YIN; MATSUMOTO, 2021; VAHIDNIA; ABBASI; ABBASS, 2020; XU, S. *et al*., 2020; ZHANG *et al*., 2017). Em sua maioria, esses estudos baseiam-se em

medidas bibliométricas, investigando relação entre informações semiestruturadas como rede de coautores, redes de cocitação e redes de palavras-chave, por exemplo. Ainda que sejam importantes abordagens para mapear e compreender a dinâmica de como a ciência e tecnologia se organizam e se relacionam, são limitadas no sentido de antecipar novidades.

Nesse sentido, apesar de estudos sobre a ciência encorajem que suas informações sejam usadas para fins de gestão, esses estudos satisfazem plenamente a necessidade de antecipação de informações para o planejamento estratégico. Um contraexemplo sobre como os estudos de fronteira de conhecimento científico não são o bastante para antecipar informações sobre tecnologia reside nas publicações do tipo '*sleeping beauty*' (bela adormecida), as quais dizem respeito a estudos de impacto, mas que demoram a ter seus conhecimentos propagados. São especialmente associados ao conhecimento tecnológico, tal que costumam ser citados primeiro em bases de patentes e depois citados por outras publicações cientificas (RAAN, 2004, 2017; WINNINK; TIJSSEN; RAAN, 2019). Nessa lógica, como muitos estudos sobre fronteira científica estão focados em compreender o trajeto do conhecimento e os atores envolvidos, não seriam capazes de extrair esse tipo de informação útil para as pesquisas de antecipação tecnológica, visando a antecipação da informação para planejamento estratégico.

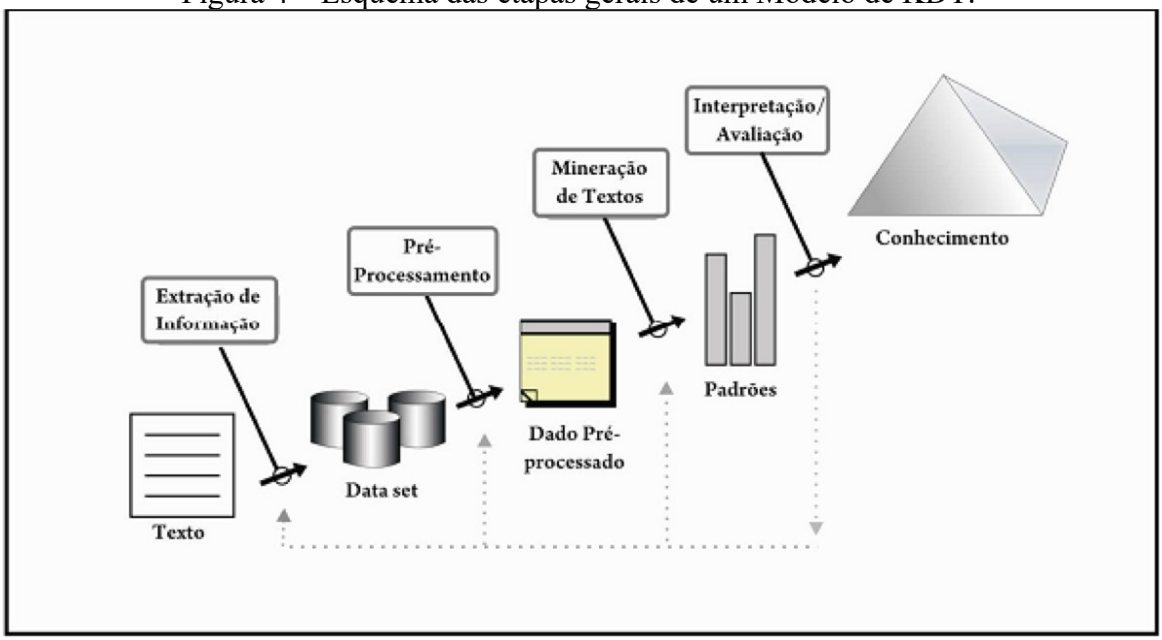
Como explicitam Efimenko e Khoroshevsky (2017), basear uma análise apenas em artigos mais citados pode ser considerado uma limitação, uma vez que sinais fracos podem aparecer em artigos (ainda) não citados, como por exemplo, nas publicações "bela adormecida" ("*sleeping beauty"*). Descobertas podem permanecer despercebidas pela comunidade científica por algum tempo, mas depois serem consideradas um avanço revolucionário (*a breakthrough*). Dessa forma, evidencia-se a necessidade de explorar o conteúdo não estruturado dessas bases, implicando em um viés de análise semântica e extração do conhecimento, para antecipar possíveis tecnologias emergentes sob a luz do conceito de sinais fracos.

2.3 DESCOBERTA DE CONHECIMENTO EM TEXTO

A Descoberta de Conhecimento em Texto (*Knowledge Discovery in Text - KDT*) tem como objetivo descobrir informações úteis (conhecimento) e padrões em base de dados de documentos. Onde um documento é compreendido como um registro textual (conjunto de códigos alfanuméricos), excluindo outras formas de dados não estruturados como registros audiovisuais (imagem, áudio e vídeo) (BUCKLAND, 1997).

O texto, como dado não estruturado, é facilmente processado e percebido por humanos, mas é significativamente mais difícil para as máquinas entenderem (ALLAHYARI *et al*., 2017). O KDT constitui-se em um processo interativo e iterativo, composto por várias etapas e refinamento, repetidos em múltiplas iterações, abarcando desde como os dados são coletados, armazenados, acessados, passando pela busca por padrões, e finalizando em como os resultados podem ser interpretados (Figura 4). Dessas, a etapa de mineração de texto é responsável pela aplicação de algoritmos com o propósito de obter informações. Caracteriza-se assim, como um campo de estudo interdisciplinar, na interseção de vários domínios relacionados, incluindo: Estatística, Computação, Recuperação de Informação (*Information Retrieval* - IR), Extração de Informação (*Information Extraction* - IE), Inteligência Artificial, Processamento de Linguagem Natural (*Natural Language Processing* - NLP), Aprendizado de Máquina (*Machine Learning* - ML) e Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases - KDD*) (FELDMAN; DAGAN, 1995; MOONEY; NAHM, 2005; GUPTA; LEAHL, 2009).

Figura 4 – Esquema das etapas gerais de um Modelo de KDT.



Fonte: Adaptado de (MOONEY; NAHM, 2005) e (GUPTA; LEAHL, 2009)

A área surge da necessidade de tornar grandes volumes de dados em algo mais compreensível e útil, de uma forma que facilite a interpretação por especialistas humanos. Com o crescente uso de computadores, houve um significativo aumento na geração de dados, por consequência um acúmulo de grandes volumes desses dados. Entretanto, esse armazenamento não garante necessariamente acesso a informações úteis contidas nesses dados. O método tradicional de transformar dados em conhecimento depende da análise e interpretação manual de especialistas. Contudo, essa forma de levantamento manual de um conjunto de dados é lenta,

cara e altamente subjetiva. Além disso, conforme os volumes de dados aumentam drasticamente, a aplicação desse tipo de análise manual torna-se impraticável em muitos domínios. Em particular, para conhecimento científico que é armazenado e propagado em textos e apresenta crescimento contínuo (BORMANN; MUTZ, 2015).

Diante do exposto, a aplicação de métodos de mineração de textos é tido como uma das principais etapas do processo de KDT, tratados inclusive, algumas vezes, como sinônimos. A mineração de texto é responsável pela descoberta de padrões ocultos nos dados extraindo automaticamente informações de recursos textuais, revelando, por fim, conhecimentos até então desconhecidos (ALLAHYARI *et al*., 2017).

Para que a tarefa de mineração de texto possa ocorrer é necessário que os textos sejam representados como um vetor numérico, para que então, os algoritmos de mineração subsequentes possam extrair informações a partir dos dados. Nesse sentido, de modo sucinto, pode-se dizer que a descoberta de conhecimento em textos conta, basicamente, com duas etapas essenciais: a primeira etapa consiste no tratamento do texto a fim de convertê-lo para uma forma estruturada e, a segunda etapa consiste na aplicação de técnicas e algoritmos de mineração para a descoberta do conhecimento.

Como alicerce da mineração de texto estão técnicas e métodos de Processamento de Linguagem Natural (NLP) e Aprendizado de Máquina (ML) (CHOWDHURY, 2003; FRADKOV, 2020; LAURIOLA; LAVELLI; AIOLLI, 2022). A NLP é uma subárea da Inteligência Artificial (AI) responsável por programar os computadores a fim de que compreendam texto ou expressem em linguagem natural. Por sua vez, o objetivo principal do ML é otimizar certas tarefas, como classificação, agrupamento, inferência e previsão, treinando as máquinas para aprenderem com dados, sem a necessidade de que todas as regras estejam explicitamente programadas, e quando expostas a novos dados, permite que os computadores as executem de forma independente.

Existem duas principais abordagens para o processamento de linguagem natural: uma abordagem estatística e outra com foco linguístico. Abordagens estatísticas se baseiam na frequência e distribuição de palavras, ou seja, quanto mais abundantes e representativos forem os dados, melhores serão os resultados. Por outro lado, não consideram a estrutura ou o significado da frase e do discurso. Em contrapartida, a abordagem linguística trata sobre a estrutura, a formação e classificação das palavras. Identificam informações sintáticas e

semânticas, bem como informações sobre o contexto do discurso (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). Os sistemas atuais de NLP com melhor desempenho usam abordagens sofisticadas de Aprendizado de Máquina, Redes Neurais (*Neural Networks* - NN) e Aprendizagem Profunda (*Deep Learning* - DL) para realizar suas análises (HIRSCHBERG; MANNING, 2015; LAURIOLA; LAVELLI; AIOLLI, 2022).

Para o KDT as técnicas de NLP atuam como tarefas secundárias auxiliares na resolução na tarefa principal de descoberta de conhecimento. Na preparação do conjunto de dados a serem minerados é recorrente aplicar na etapa de pré-processamento análises morfológicas (*lemmatization, stemming, part-of-speech tagging,* etc), análises sintáticas (*parsing*), análises semânticas lexicais (*Named Entity Recognition - NER*, *Word Sense Disambiguation* - WSD*, Sentiment Analysis - SA*) entre outras; ou mesmo basear a etapa de mineração em estatísticas como frequência do termo (*Term Frequency - TF*) e frequência no documento (*Document Frequency* - DF) (ALLAHYARI *et al*., 2017).

Sobre o aprendizado de máquina, pode ser dividido, principalmente, em supervisionado ou não-supervisionado (DARGAN *et al.*, 2020). Caracteriza-se como supervisionado quando é oferecido ao modelo um conjunto de dados de treino com os determinados rótulos esperados. Nesse sentido, o modelo produzido aprende a identificar e replicar um padrão conhecido. Por outro lado, o aprendizado não-supervisionado é voltado a desvendar a estrutura implícita no conjunto de dados. Não se sabe a priori o que se está procurando e, neste sentido, são utilizados algoritmos para organizar e apresentar o resultado obtido. A descoberta de conhecimento em texto se concentra na descoberta de propriedades (previamente) desconhecidas nos documentos. A complementar, o uso de aprendizado de máquina se destaca na mineração de texto pelos modelos para representações vetoriais. Sob essa óptica, o aprendizado de máquina e a mineração de texto são áreas de estudo que se sobrepõem significativamente. Em outras palavras, quando técnicas de ML evoluem, a capacidade dos modelos de mineração de texto também progridem.

Identificar as informações relevantes sobre dados não estruturados de alta dimensão como um texto é um grande desafio, sendo necessários algoritmos adequados para encontrar ou gerar recursos interpretáveis que possam contribuir de forma significativa durante a análise dos dados (BERNARD; LEBBOSS, 2017). Um dos desafios da mineração de texto surge da alta dimensionalidade associada à linguagem natural, onde cada palavra do texto é considerada uma variável representando uma dimensão.

De início, métodos mais simples baseados em classificadores lineares, como *Support Vector Machines* (SVM), eram mais implementados em comparação às redes neurais (NN). No entanto, a partir de 2010, o uso de processamento por GPU (*Graphic Processing Unit*) renovou a aptidão das redes neurais. Contudo, os modelos ressurgem implementando *multi-layers* e o algoritmo *backpropagation,* sob o nome de *Deep Learning* (DL), oferecendo resultados superiores e capazes de comportar problemas mais complexos. Nesse momento, modelos de redes neurais *Deep Feedforward* e *Recurrent Neural Networks* (RNN) se destacavam nas tarefas de reconhecimento de padrões e aprendizado de máquina. À luz da perspectiva do NLP, os modelos LSTM (*Long Short-Term Memory*) de RNNs e a abordagem de *embedding Word2Vec* (CBOW ou *Skip-gram*) se mostraram particularmente eficientes (FRADKOV, 2020; LAURIOLA; LAVELLI; AIOLLI, 2022).

Todavia, o ano de 2018 foi um ano de ruptura, visto que o uso de LSTM e RNNs para problemas de NLP foi ultrapassado devido ao conceito de *Transfer Learning* com avanço de técnicas de *Deep Learning* baseadas na arquitetura de *Transformers* em conjunto com *Attention Mechanisms* (VASWANI *et al.*, 2017; WOLF *et al.*, 2020). O *Transfer Learning* é um conceito para se referir ao reaproveitamento de parâmetros de um modelo já treinado de modo que sejam usados em outro *corpus* e/ou para outra tarefa, dando origem a uma nova classe de modelos conhecidos por LLMs - *Large Language Models* (ZHAO *et al.*, 2023). A título de curiosidade, o algoritmo GPT, criado pelo centro de pesquisa OpenIA®, é um *Autoregressive Language Model* (*decoder only*) e o algoritmo BERT criado pela Google®, é um *Autoencoding Language Model* (*encoder only*), ambos fundamentados na arquitetura *transformer* e *self-attention mechanism.*

A seguir é detalhado o processo de representação vetorial das palavras, sendo aprofundado o modelo de linguagem pré-treinado BERT (DEVLIN *et al.*, 2019).

2.3.1 Representação Vetorial das Palavras

Todo processamento computacional necessita de uma representação numérica para realizar seus cálculos. Salton, Wong e Yang (1975) descrevem o Modelo de Espaço Vetorial (*Vector Space Model* - VSM) para representar textos e documentos na forma de vetores, permitindo que operações algébricas possam ser realizadas. A extração de atributos (*feature extraction*) ou vetorização (*text vectorization*), é o nome atribuído a este processo de transformar o texto (dado não estruturado) em uma forma estruturada de armazenar informações

(LIANG *et al.*, 2017). Estes vetores, por sua vez, podem ser utilizados como entrada para diferentes tipos de tarefas de processamento de linguagem natural (NLP), como, similaridade entre palavras, recuperação de informação, classificação de documentos, análise de sentimentos, entre outras.

One-hot encoding (OHE) e *Word Embedding* (WEs) são, atualmente, as principais técnicas para representação vetorial em NLP. A primeira obtém o vetor por análise estatística, baseada na frequência das palavras. A segunda, obtém por análise semântica, baseada na funcionalidade das palavras nos textos. Para gerar tais vetores, diferentes técnicas podem ser utilizadas, as estatísticas como BOW[17] (HARRIS, 1954) e TF-IDF[18] (ROBERTSON, 2004) não consideram a ordem ou estrutura nas palavras no texto e geram vetores esparsos. Por sua vez, técnicas de *word embeddings*, permitem que as palavras de um conjunto de texto sejam mapeadas em vetores densos, considerando as relações entre as palavras (sintática e semântica) em sua representação numérica vetorial. Isso viabiliza a análise quantitativa de características qualitativas. No caso, a característica qualitativa de interesse é a semântica das palavras que se objetiva representar numericamente estas relações entre as palavras de modo que incorpore seu significado contextual. Isto possibilita que os vetores das palavras sejam análogos ao significado da palavra.

Os WEs são aprendidos através do treinamento de redes neurais (*Deep Learning*), são exemplos os modelos *Word2vec* (MIKOLOV *et al.*, 2013) e BERT (*Bidirectional Encoder Representations from Transformers)* (DEVLIN *et al.*, 2019). No entanto, a semântica representada pelo *embedding* pode se tratar de uma semântica estática ou contextual. Abordagens como *Word2vec* (C-BOW ou *Skip-gram*), GloVe (PENNINGTON; SOCHER; MANNING, 2014), entre outros, são modelos que fornecem *embeddings* estáticos, pois exibem um único vetor no espaço semântico, ou seja, polissemia[19] e homonímia[20] não são tratadas adequadamente. Abordagens, mais recentes, como ELMo (PETERS *et al*., 2018), UMLFiT (HOWARD; RUDER, 2018), BERT (DEVLIN *et al.*, 2019), GPT-2 (RADFORD *et al.*, 2019), GPT-3 (BROWN *et al.*, 2020) para citar os mais conhecidos, fornecem *embeddings* com semântica dinâmica, contextualizada, isto é, são capazes de gerar *embeddings* diferentes para uma mesma palavra que apresente contextos distintos. Um progresso significativo para o

[17] *Bag-of-Words*
[18] *Term Frequency-Inverse Document Frequency*
[19] Uma palavra que reúne vários significados.
[20] São duas ou mais palavras que possuem a mesma grafia ou a mesma pronúncia, mas com significados diferentes entre si.

processamento de linguagem natural considerando que palavras frequentes, de uso corrente, tendem a apresentar mais sentidos, conforme o Princípio da Versatilidade Econômica das Palavras (ZIPF, 1949).

Atualmente, para obter os *embeddings* modelos de linguagem *deep learning* pré-treinados que apresentam arquitetura com base em '*transformer models' e 'attention mechanisms'* (VASWANI *et al.*, 2017) como BERT (DEVLIN *et al.*, 2019) e GPT-3 (BROWN *et al.*, 2020), representam o estado da arte em NLP. E, substituíram técnicas precedentes de modelos sequenciais, como, LSTM (*Long Short-Term Memory*), RNN (*Recurrent Neural Network*), GRU (*Gated Recurrent Unit*), entre outros, por apresentarem resultados superiores.

Vale ressaltar também que modelos baseados no BERT configuram uma classe de modelos (MOHAMMED; ALI, 2021), entre eles, RoBERTa; DistilBERT, BART, BigBird, ALBERT, DynaBERT, ConvBERT e vários outros - todos baseados na estrutura *Bidirectional Encoder Representations from Transformers* apresentado pelo primeiro modelo BERT (DEVLIN *et al.*, 2019). Esses modelos da família BERT apresentam ajustes específicos que podem atender necessidades pontuais dependendo do problema. A justificativa da escolha do modelo BERT para esta pesquisa pode ser encontrada adiante na seção 4.2.2. A seguir é detalhado como obter o *embedding* de uma palavra com o modelo BERT.

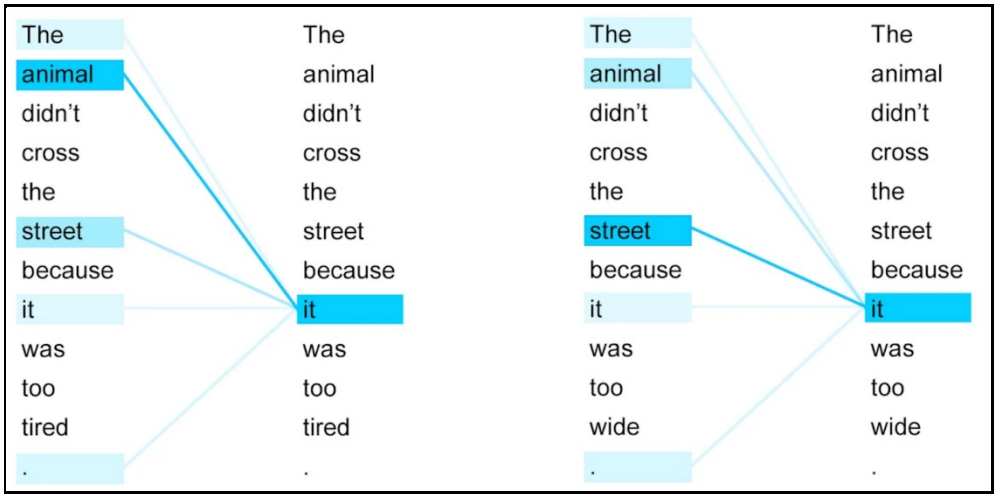
2.3.1.1 BERT

O BERT® (***B**idirectional **E**ncoder **R**epresentations from **T**ransformers*), proposto por Devlin *et al*. (2019), é um modelo para NLP pré-treinado em um grande corpo de texto (*Wikipedia*® e *BooksCorpus*®) executando duas tarefas: linguagem mascarada (*Masked Language Modelling* - MLM) e previsão da próxima sentença (*Next Sentence Prediction* - NSP). A tarefa de linguagem mascarada tem por objetivo fazer com que o modelo preveja, com base no contexto da sentença, qual palavra foi ocultada (mascarada, "*masked*"). O objetivo do treinamento de previsão da próxima sentença é avaliar a capacidade do modelo em entender o significado geral de uma sentença ao invés apenas das palavras específicas. Com essa finalidade o modelo realiza a previsão se dadas duas sentenças elas apresentam uma conexão lógica e sequencial ou se sua relação é simplesmente aleatória. Este tipo de treinamento permite que o modelo aprenda uma representação semântica do texto sem a necessidade de dados rotulados. Depois que esse pré-treinamento é realizado uma vez, pode-se reutilizar com eficiência essas representações geradas pelo BERT para diferentes tarefas.

Historicamente, modelos de linguagem conseguiam lidar apenas com texto de maneira sequencial. O modelo BERT se diferencia por sua arquitetura baseada em *Transformers* e *self-attention mechanism*, permitindo que o modelo analise o texto simultaneamente nos dois sentidos. Ou seja, para representar uma palavra em uma sentença, BERT usa tanto o contexto da palavra anterior quanto da palavra seguinte. A arquitetura *Transformer* permite que o modelo BERT compreenda o contexto da palavra, executando constantes etapas de codificação (*encoder*) empregando o mecanismo de atenção (*attention mechanism*) (VASWANI *et al.*, 2017) decidindo em cada etapa quais partes da sequência são importantes. Análogo a interação humana ao ler um texto, em que enquanto se lê uma palavra em particular mantêm-se em mente palavras-chave anteriores relevantes para compreensão do texto.

Isso permite que o modelo BERT, por exemplo, como mostra o exemplo na Figura 5, compreenda em cada caso a qual palavra um mesmo pronome se referencia. Isto é, no primeiro exemplo o pronome *'it'* se referencia ao substantivo 'animal' e no segundo caso *'it'* se referencia ao substantivo 'rua'. Assim, por sua capacidade de *embedding* contextual dinâmico o BERT é capaz de prover representações vetoriais distintas para uma mesma palavra '*it'* devido a distinção de contexto em suas ocorrências.

Figura 5 – Distribuição de *self-attention* do encoder para a palavra *'it'* de um *transformer* treinado.



Fonte: Adaptado de Google AI Blog[21] (Uszkoreit, 2017)

O modelo BERT (DEVLIN *et al.*, 2019), apresenta duas formas de uso, a primeira refere-se a adicionar uma camada final de ajuste fino (*fine-tunning*) ao modelo de base para execução de tarefas específicas de acordo com a necessidade do estudo. Isso possibilita que

[21] *Transformer: A Novel Neural Network Architecture for Language Understanding* - https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

parâmetros de um modelo já treinado sejam aplicados a outro *corpus* e outra tarefa. Desta forma, ao invés de treinar os modelos do zero, o conhecimento incorporado em um modelo de aprendizado de máquina pré-treinado é reutilizado e ajustado para uma outra tarefa e domínio específico. Essa reutilização de *embeddings* aprendidos em outras tarefas é chamado de *transfer learning.* A segunda, tem a finalidade de extração de atributos (*feature-based approach*), ou seja, gerar *embeddings* de palavras contextualizadas, sem a necessidade de ajustes nas tarefas do modelo BERT. Isso significa que para a extração de atributos, o modelo aprende os *embeddings* dos termos ao executar as mesmas tarefas de seu pré-treinamento (MLM e NSP) em um dado corpus de domínio específico.

O *embedding* da palavra no modelo BERT ocorre pelo contexto da frase na qual ela se encontra como visto na Figura 5. Por essa razão, é capaz de obter *embeddings* dinâmicos, isso significa a capacidade de gerar *embeddings* diferentes para uma mesma palavra, considerando a frequência com a qual ela ocorre em frases de contextos distintos. Como, por exemplo, a palavra 'letra', pode ocorrer em contextos distintos referindo-se ao elemento básico do alfabeto, a caligrafia de um indivíduo ou o texto de uma canção. Uma abordagem de representação estática como *Word2vec* apresentaria um único *embedding* para a palavra 'letra', mas o BERT é capaz de apresentar um vetor distinto para cada um desses sentidos de uma mesma palavra.

Além disso, outra vantagem do modelo BERT é sua habilidade de lidar com palavras fora do vocabulário (*Out of Vocabulary* - OOV). Em NLP esse problema de OOV se configura quando ocorrem palavras que não estão no conjunto de treinamento, mas aparecem no conjunto de teste ou nos dados reais (KHODAK *et al.*, 2018; HU *et al.*, 2019). Essas palavras desconhecidas podem estar relacionadas, por exemplo, a termos técnicos de domínios de conhecimento específico ou, também, textos informais de internet, onde ocorrem gírias, neologismos e erros de digitação que podem ser consideradas palavras desconhecidas. A habilidade do modelo BERT em lidar com palavras OOV se dá devido à implementação do tipo de algoritmo de *token*ização, *sub-word tokenizer,* contido no modelo pré-treinado BERT. Originalmente, o modelo BERT utiliza do tokenizador *WordPieces*® (WU *et al.*, 2016), mas existem outros como o *Byte-Pair Encoding*® (BPE) (SENNRICH *et al*., 2015), por exemplo.

O processo de tokenização tem como objetivo representar o texto original (entrada não estruturada) em unidades menores chamadas *tokens*, elementos discretos adequados para o aprendizado de máquina. Um *token* é uma sequência de caracteres que representa uma unidade de significado em um texto. Essas unidades de significado podem ser as próprias palavras,

subpalavras ou caracteres individuais, dependendo do algoritmo de tokenização implementado. A tokenização em subpalavras, como ocorre no modelo BERT, baseia-se no princípio de que palavras usadas com frequência não devem ser segmentadas em subpalavras, mas palavras menos frequentes ou que apresente afixos comuns devem ser decompostas em subpalavras significativas para beneficiar seu processamento. Este tipo de tokenização permite lidar com palavras OOV e, consequentemente, os *embeddings* obtidos por meio de modelos BERT são capazes de lidar com palavras OOV e palavras raras (palavras que ocorrem com baixa frequência).

O modelo de linguagem pré-treinado BERT pode ser encontrado em mais de uma construção: modelo *base* ou *large* e *case* ou *uncased*, isto é, uniformizando letras maiúsculas e minúsculas. O modelo BERT *base* usa 12 (doze) camadas de *transformers encoders* e o *large*, o dobro, 24 (vinte e quatro) (DEVLIN *et al.,* 2019). Para obter os *embeddings* a rede neural é treinada nas duas tarefas supracitadas e, a partir disso, os pesos das camadas ocultas são estimados. Para o modelo BERT *base*, a saída para cada *token* são 12 camadas de vetores que podem ser usadas como *embedding*. Dessa forma, para obter vetores individuais para cada um dos *tokens*, diferentes estratégias podem ser adotadas, tais como *pooling strategies* e *layer choice*, ou seja, como combinar (somar, ponderar ou concatenar) as camadas selecionadas.

Em seu artigo Devlin *et al*. (2019) mostram que a concatenação das quatro últimas camadas foi a melhor estratégia para extração de atributos dos *tokens* considerando a tarefa de NER (*Name Entity Recognition*). Entretanto, as demais estratégias eram tão boas quanto, sendo aconselhável que dependendo da tarefa final, diferentes estratégias de seleção e combinação das camadas sejam testadas a fim de obter a melhor representação. Jawahar, Sagot e Seddah (2019) mostraram que as primeiras camadas do BERT capturam informações ao nível da frase, enquanto as camadas intermediárias captam informações sintáticas, e as camadas finais informações semânticas. Ou seja, à medida que os *embeddings* se aprofundam nas camadas da rede, eles representam mais informações contextuais, assim, a combinação de camadas fornece uma melhor representação (WANG; KUO, 2020). Por essa razão as estratégias ficam em torno da combinação das 4 (quatro) últimas camadas.

2.3.2 Análise de Redes Complexas

A representação de informações por redes proporciona um modo sistemático de análise visual, possibilitando revelar estruturas e conhecimentos ocultos que seriam difíceis de reconhecer por outros meios (NEWMAN, 2003; LIU *et al*., 2018). Nesse sentido, é uma das ferramentas que mais cresce a implementação no intuito de descoberta de conhecimento.

A análise de redes pode ser vista como um desdobramento da Teoria dos Grafos, uma área da matemática discreta, sendo a linguagem matemática adotada para mensurar as propriedades das redes (NEWMAN, 2003). O problema das pontes de *Königsberg* é tido como marco inicial da teoria de redes (GRIBKOVSKAIA; HALSKAU; LAPORTE, 2007). Nesse problema matemático clássico, Leonard Euler em 1736, usou um esquema em grafo para representar o conjunto de pontes da pequena cidade de *Königsberg*, na Prússia. Dessa forma a Teoria dos Grafos é pilar para descrever as propriedades das redes, tanto na forma gráfica (visual), quanto na forma matricial (matemática) (NEWMAN, 2003; BOCCALETTI *et al.*, 2006).

Assim, um grafo é uma estrutura usada para representar 'coisas' e suas relações, sendo composto de dois componentes - o conjunto de nós (também chamados de vértices ou nodos) e o conjunto de arestas (também chamados de arcos ou conexões). Cada aresta conecta um par de nós, indicando que há uma relação entre eles. Essas arestas podem ser direcionadas ou não-direcionadas, dependendo da reciprocidade da relação, e também podem apresentar um valor associado a esta conexão relativo à intensidade da relação (peso) (BOCCALETTI *et al.*, 2006). Essa estrutura relacional estabelecida em grafo é base para a análise de redes, explorando sua estrutura e identificando padrões.

Destarte, uma rede pode ser entendida como um conjunto de elementos interconectados (NEWMAN, 2003), e são comumente utilizadas como abstrações das relações presentes na natureza (quase tudo pode ser representado em uma estrutura de rede) (BARABÁSI; BONABEAU, 2003). Com o objetivo de entender como emergem e evoluem as redes reais, ou seja, as redes que se formam na natureza, tecnologia e sociedade (NEWMAN, 2003; BARABÁSI; BONABEAU, 2003), o campo de estudo de ciência de redes (*Network Science*) tem como premissa que, apesar das diferenças aparentes entre essas redes, elas satisfazem um conjunto fundamental de leis e mecanismos que podem ser interpretados a partir de um conjunto unificado de ferramentas e princípios (BARABÁSI, 2013).

Embora a Teoria de Grafos seja uma área sedimentada de estudos matemáticos e o conceito da análise de redes seja também teoricamente bem estabelecida. A análise de redes complexas (redes reais) como se conhece hoje, despontou no final dos anos de 1990 em decorrência dos estudos de Duncan J. Watts e Steven Strongatz (1998) e Albert-Lászlő Barabási e Réka Albert (1999) enfatizando as características presentes nas redes reais. Esses trabalhos se destacam por apresentarem os conceitos de *Small-World Networks* (pequeno mundo) (WATTS; STROGATZ, 1998) e *Scale-free Networks* (escala-livre) (BARABÁSI; ALBERT, 1999), respectivamente, mostrando que nas redes reais, diferentemente da teoria tradicional (ERDÖS; RÉNYI, 1959), a distância média entre os nós permanece pequena, mesmo quando a rede se torna muito grande (WATTS; STROGATZ, 1998). Assim como, a maioria dos nós apresenta um número de conexões muito baixo, ao mesmo tempo que, por outro lado, alguns nós são altamente conectados (*hubs*) (BARABÁSI; ALBERT, 1999). Essas características podem ser encontradas especialmente ao se analisar redes sociais. A partir de 2000, Girvan e Newman, (2002) exploraram estruturas sociais por meio do uso de redes e da teoria dos grafos, contribuindo para a difusão da técnica de análise de redes. Por essa razão, muitas pesquisas que fazem uso de análise de redes complexas a nomeiam como análise de redes sociais (*Social Network Analysis* - SNA).

Diante desse cenário, a análise de redes reais, ou redes complexas, é uma área relativamente recente. Pode-se citar o ano de 2005 como um marco, quando o Conselho Nacional de Pesquisa dos EUA definiu a Ciência de Redes como um novo campo de pesquisa básica (MOLONTAY; NAGY, 2019). A área percorreu um longo caminho para estabelecer uma base comum, chegar a um acordo sobre definições e conciliar abordagens adotadas por campos tão díspares quanto ciências sociais, física, biologia, ciência da computação e matemática aplicada. Contudo, apesar das dificuldades inerentes a campos interdisciplinares, a área de Ciência de Redes vem avançando, se disseminando e se consolidando (VESPIGNANI, 2018; MOLONTAY; NAGY, 2019).

Como visto, as redes podem se referir à representação de diferentes elementos para diferentes domínios. Levando em conta os objetivos da pesquisa proposta nesta tese, é dada atenção especial aos estudos dedicados a representação do conhecimento presente em bases textuais. Não são raros os estudos que tem as bases de publicações científicas como fonte para estabelecer redes de palavras. Estudos com esse propósito recebem o nome de *co-word networks* ou *co-word analysis* e a visualização das relações entre palavras na forma de rede é destacada por fornecer entendimento intuitivo (KATSURAI; ONO, 2019).

As redes de palavras são geradas conectando pares de palavras com base em critérios que definam a coocorrência, ou seja, a presença pareada dentro de uma unidade de texto especificada, como uma frase, parágrafo ou documento. A representação das relações entre palavras configuram um grafo não-direcionado, em razão da reciprocidade na relação. O pretexto dessa análise baseia-se que quando termos ocorrem com certa frequência, próximos uns aos outros, estes têm uma relação mais forte do que termos que ocorrem com menos frequência. Pares de palavras que coocorrem podem ser chamadas de "vizinhos". Dessa forma, os termos que coocorrem tendem a expressar a existência de temáticas recorrentes, assuntos centrais e conceitos que constituem e estruturam uma área de estudo investigada (LI; ZHANG; HONG, 2020).

De mais a mais, redes reais quase nunca apresentam características estáticas, sendo que muitas redes demostram um comportamento de crescimento ao longo do tempo. No entanto, o crescimento não é a única maneira pela qual as redes evoluem; nós e arestas podem surgir e desaparecer durante a vida útil de uma rede, por exemplo. Sob essa óptica, a análise temporal de redes permite explorar uma nova dimensão de processos dinâmicos em redes. A combinação da complexidade da estrutura da rede com a complexidade induzida pelo comportamento temporal, abrem uma nova gama de métricas e conhecimentos que podem ser descobertos onde frações de conhecimento antigo e novo estão interligados. Embora conceitualmente compreensível, a análise temporal de redes é ainda uma área de pesquisa em desenvolvimento devido à complexidade computacional envolvida para a operacionalização de suas análises (VEGA; MAGNANI, 2018). Dentre os aspectos que podem ser explorados em redes temporais, dois se destacam pela importância nesta pesquisa, a análise de formação de comunidades (sub-grafos) e análise de ruptura (*Burst Analysis*).

A análise de comunidades não se restringe às redes temporais, sendo uma análise topológica da rede, em que uma comunidade pode ser compreendida como a formação de um subgrafo na rede. Entretanto, com a dimensão temporal, podem-se investigar as mudanças na estrutura da rede (YANG; LESKOVEC, 2015). A estrutura de comunidades é também uma das características que se destacam no estudo de redes complexas, ou redes reais. Diz-se que uma rede tem estrutura de comunidade se existe a ocorrência de um conjunto de nós, de modo que cada conjunto de nós seja densamente conectado internamente em comparação com o restante da rede (YANG; LIU; LIU, 2010). O conceito de densidade descreve o nível de ligação entre os nós de um grafo, tal que quanto mais conexões diretas existem entre os nós, mais denso se torna o grafo, ou subgrafo.

A investigação das comunidades permite uma compreensão complementar da rede, uma vez que, as comunidades podem ser vistas como meta-nós dentro da rede (NEWMAN; GIRVAN, 2004; NEWMAN, 2006). Além disso, internamente uma comunidade pode apresentar um comportamento diferente quando comparado ao da rede como um todo. Todavia, encontrar subgrafos dentro de uma rede temporal é uma tarefa computacionalmente difícil. O número de subgrafos, quando existem, é normalmente desconhecido e frequentemente apresentam tamanhos e/ou densidades desiguais, pois são determinados organicamente pela própria estrutura topológica da rede. Considerando o desafio da tarefa, são extensivos os estudos e métodos propostos na literatura a encontrar subgrafos temporais com diferentes níveis de sucesso (ARAUJO *et al*. 2016; DONDI; HOSSEEINZADEH, 2021; GAO *et al.*, 2020; LI, Y. *et al.*, 2021; MA *et al.*, 2020; MORIANO; FINKE; AHN, 2019; U; GE; WU, 2020; ROZENSHTEIN *et al.*, 2020; WANG *et al.*, 2022; ZHANG *et al*., 2022; ZHU C., 2022; YANG *et al.,* 2023).

Por fim, identificar subgrafos que apresentem um comportamento de ruptura (*burst*) em um período de tempo configura uma classe particular de problema. Tendo como objetivo buscar a ocorrência de mudanças emergentes em redes temporais (QIN *et al*. 2019). A escolha dessa abordagem para monitoramento da rede de palavras é apresentado na seção 4.2.2. Por ora, esse tema é aprofundado na seção a seguir.

2.3.2.1 Análise Burst

O conceito da métrica "*burst*" não possui uma tradução literal para o português, mas está associado com o conceito de ruptura; uma interrupção de continuidade, seja de um crescimento notável em um curto período de tempo ou surgimento/desaparecimento repentino de nós ou conexões. Kleinberg (2003) é frequentemente referenciado como marco dos estudos de *burstness* tendo proposto o algoritmo mais tradicional para identificação desse tipo de fenômeno. As contribuições de seus trabalhos forneceram uma compreensão mais profunda do comportamento *burst* e suas implicações para a ciência de redes, servindo como base para pesquisas subsequentes investigando padrões temporais ocultos e a dinâmica de sistemas complexos.

Em redes reais complexas, o comportamento de ruptura (*burst*) é comum na evolução da rede, onde períodos de alta atividade são seguidos por relativa inatividade, visto como um reflexo da dinâmica temporal e comportamento de atividades humanas. Nesse sentido,

atualmente, o avanço computacional oportuniza a investigação de redes temporais, de modo que se apresenta como um tema em destaque. Pesquisas recentes se dedicam a métrica *burstness* para investigar ao longo do tempo uma alteração repentina e/ou inesperada, bem como um comportamento de crescimento explosivo em redes complexas dinâmicas (CHU *et al*., 2019; DAI *et al.*, 2021; LI *et al*., 2022; QIN *et al*., 2019; QIN *et al.*, 2022). Esses algoritmos incorporaram métodos estatísticos, análise de séries temporais ou abordagens de aprendizado de máquina para identificar um comportamento *burst* nos dados das redes temporais. Permitindo focos de investigação em identificar padrões *bursts* segundo critérios de proporção, duração ou intensidade, a depender da necessidade de compreensão do padrão para cada estudo.

Contudo, se por um lado encontrar as partes que mudam mais rapidamente é uma tarefa de interesse central na análise de redes temporais (CHU *et al.*, 2019), por outro lado, informações referentes a este tipo de comportamento *burst* são de difícil obtenção por estarem ocultas em outras dinâmicas complexas da rede real (ZHAO *et al.,* 2019). Pode-se mencionar como principais desafios fatores como ruído e variabilidade nos dados e problemas de escalabilidade. Em outras palavras, diferenciar um comportamento *burst* genuíno de uma flutuação na variabilidade dos dados é uma dificuldade, além disso, essa dinâmica *burst* é considerada usualmente ruído por modelos de aprendizado de máquina e, portanto, uma informação muitas vezes ignorada (CHANDOLA; BANERJEE; KUMAR, 2009). Também, como mencionado anteriormente, redes reais complexas são de difícil processamento, pois podem conter nós e arestas na ordem de milhões, tal que analisar esses conjuntos massivos de dados não é uma tarefa computacionalmente trivial.

Associado ao objetivo da pesquisa apresentado por essa tese, Dernis, Squicciarini e Pinho (2016), Li e Chu (2016) e Katsurai e Ono (2019) são alguns autores que propuseram modelos que implementam o conceito de *burst* para detectar, na devida ordem, o surgimento de tecnologias e para a fronteira da ciência e tópicos de pesquisa *burst*, ou seja, tópicos que têm crescido rapidamente, em vez de tópicos simplesmente populares. Sobretudo, como explica Chu *et al.* (2019), uma vez que um subgrafo (comunidade), pode acumular lentamente sua densidade por um longo período de tempo, nem todo subgrafo denso pode ser considerado um subgrafo de comportamento de ruptura (*Density Bursting Subgraph* - DBS). Portanto, os métodos de detecção de subgrafos densos (temporais) existentes não podem ser estendidos diretamente para detectar subgrafos de densidade com comportamento de ruptura (*burst*).

Um DBS não é necessariamente o subgafo de maior densidade, mas é um subgrafo que acumula sua densidade na velocidade mais rápida durante um intervalo de tempo. Aprofunda-se a compreensão desse conceito por meio do algoritmo Top-k DBS proposto por Chu *et al.* (2019), no qual a densidade de um subgrafo é medida pela coesão (*cohesiveness*), Equação (1), que é a força de conexão média entre os nós do subgrafo, e a velocidade de acúmulo de densidade é medida por *burstiness*, Equação (2), ou seja, a razão entre o ganho de densidade de um subgrafo e o tempo para acumular este ganho (CHU *et al.,* 2019).

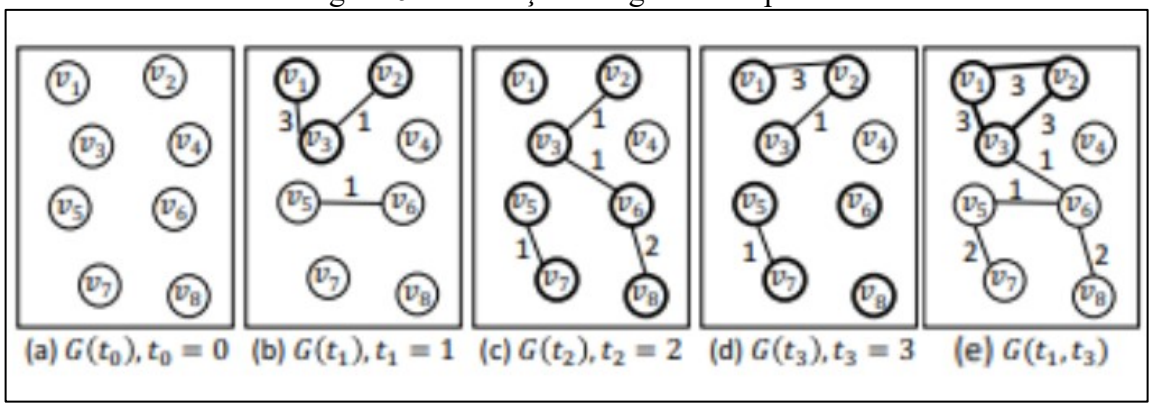
$$q_x(t_h) = X^T A(t_h)\, X \tag{1}$$

Onde, $q_x(t_h)$ é a função *cohesiveness* para um subconjunto de nós x no instante de tempo t_h e $A(t_h)$ é a matriz de afinidade no tempo t_h e X e X^T são o vetor e o vetor transposto com pesos positivos que indicam a importância do nó no subgrafo.

$$g(x,T) = \frac{\sum_{h=b+1}^{e} q_x(t_h)}{t_e - t_b} = \frac{X^T A(t_{b+1}, t_e) X}{t_e - t_b} \tag{2}$$

Onde, $g(x,T)$ é a função *burstness* para um subgrafo induzido pelos nós x no período $T = (t_b, t_e]$, definido pela densidade calculado pela Equação 2 dividido pela duração do intervalo de tempo.

Figura 6 – Ilustração dos grafos temporais.



Fonte: Extraído de Chu *et al.* (2019)

Para compreender mais a fundo e interpretar as equações detalha-se como Chu *et al.* (2019) elaboraram as notações. Para os autores, um grafo temporal (Figura 6), representado por $G(t_0, t_c) = \{G(t_0), G(t_1), \dots, G(t_c)\}$, é uma sequência dos instantes (*snapshots*) que ocorreram nos tempos $t_0, t_1, \dots, t_c$, respectivamente. Cada instante (*snapshot*) é um grafo estático. O grafo temporal $G(t_0, t_c)$ é inicializado como um grafo vazio no tempo t_0, ou seja, o instante $G(t_0)$

contém apenas o conjunto de vértices *V*, sem arestas (conexão). As mudanças ao longo do tempo acontecem nos pesos das arestas. Os instantes $G(t_1), \ldots, G(t_c)$ carregam as atualizações nos pesos das arestas nos instantes de tempo $t_1, \ldots, t_c$, respectivamente. O instante que ocorre no tempo t_h pertencente ao conjunto de tempos $\{t_1, \ldots, t_c\}$ é um grafo estático denotado por $G(t_h) = (V, A(t_h))$, onde t_h é o tempo, e $A(t_h)$ é a matriz de afinidade que define as atualizações nos pesos das arestas no tempo t_h, que é central para os cálculos Equação 1 e Equação 2. Pontua-se que uma matriz de afinidade é como uma matriz de adjacência, comumente utilizada para representação das redes, exceto que o valor de um par de pontos expressa o quão semelhantes esses pontos são. Dessa forma, a afinidade age como os pesos das arestas no grafo.

Para cada instante $G(t_h)$, a matriz de afinidade é representada por $A(t_h)$ *n* x *n* não-negativa, onde *n* é o número de vértices em *V* e cada entrada $a_{ij}(t_h)$ na *i*-éssima linha e *j*-éssima coluna de $A(t_h)$ é a atualização do peso na aresta entre o *i*-éssimo vértice $v_i \in V$ e o *j*-éssimo vértice $v_j \in V$. Existe um vértice entre v_i e v_j no instante $G(t_h)$ se, e somente se, $a_{ij}(t_h) > 0$. O intervalo de tempo, representado por $T = (t_b, t_e] = \{t_{b+1}, t_{b+2}, \ldots, t_e\}$, é um conjunto de instantes entre tempo inicial, t_b (*begin*) e o tempo final, t_e (*end*), excluindo t_b. A duração de $T = (t_b, t_e]$ é $t_e - t_b$. Um grafo acumulado durante $T = (t_b, t_e]$ é um grafo estático $G(t_{b+1}, t_e) = (V, A(t_{b+1,}t_e))$, onde $A(t_{b+1,}t_e) = \sum_{h=b+1}^{e} A(t_h)$ é a matriz de afinidade acumulada (Figura 6 (e)).

Um subgrafo temporal é uma sequência de subgrafos que são induzidos por um conjunto de vértices ponderados (V_x pertencente ao conjunto *V* de vértices do grafo) nos instantes do intervalo $G(t_{b+1}, t_e)$. A cada vértice v_i é atribuído um peso positivo x_i, tal que $\sum x_i = 1$ para $x_i > 0$ e todos os outros vértices que não pertencem ao conjunto de vértices do subgrafo *S* são atribuídos zeros (0). Dessa forma, os subgrafos podem ser representados pela tupla (x, T). Então, $G_x(t_h)$ é um subgrafo induzido pelos vértices V_x no instante $G(t_h)$ no tempo t_h. No exemplo, para $T = (t_0, t_3]$ o subgrafo temporal (x, T) é o subgrafo induzido pelo conjunto de vértices $\{v_1, v_2, v_3\}$ e $x = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0,0,0,0,0\right\}$ a partir dos *snapshots* $G(t_1)\ G(t_2)\ e\ G(t_3)$.

O subgrafo DBS (x^*, T^*) é um máximo local do problema de programação linear *Mixed Integer Programming* (MIP) (Equação 3):

$$\text{MIP} = \underset{(x,T)}{\operatorname{argmax}}\, g(x,T) \quad s.a\ x \in \Delta^n, T = (t_b, t_e], t_e - t_b \geq \theta \tag{3}$$

Isso significa que para encontrar um DBS, procura-se pelo máximo local do MIP atualizando iterativamente x e T aumentando monotonamente $g(x,T)$ (*burstness* (Equação 2)). Para atualizar x resolve-se o problema CQP (*Constained Quadratic Programming*) e para atualizar T resolve-se o problema MDS (*Maximum Density Segment*). Assim, dado uma sequência temporal de grafos pode-se identificar o conjunto de nós que apresenta o comportamento de crescimento mais rápido em suas conexões, estabelecendo um subgrafo de comportamento *burst.*

2.4 TERMINOLOGIA CIENTÍFICA

A ciência é uma ação social e, portanto, uma importante habilidade é a de comunicação viabilizada pela linguagem (HALLIDAY, 2004). Todavia, compreender o vocabulário e os conceitos apresentados pela ciência pode ser desafiador (BARAM-TSABARI *et al.*, 2020; SIEGFRIED, 2020). Ao desenvolver seus trabalhos, pesquisadores muitas vezes realizam descobertas ou criam novos artefatos e conceitos, tangíveis ou intangíveis, e são compelidos a nomeá-los (PONTES, 1997). Podem ser palavras únicas completamente novas, podem ser acrônimos, podem ser locuções ou termos compostos, ou ainda, atribuir novo significado a um termo comum já existente (ISO 704, 2009).

Esses novos termos podem dizer respeito às descobertas na natureza, como partículas subatômicas (*quarks*); sobre criações, como materiais, metamateriais[22] e elementos químicos (Teflon® - *politetrafluoretileno* (PTFE)); técnicas (VoIP- *Voice over Internet Protocol*); instrumentos ou equipamentos (laser - *light amplification by stimulated emission of radiation*) e outros. Termos técnicos muitas vezes são cunhados pela necessidade de ser preciso na descrição de algumas descobertas (PONTES, 1997; HIRST, 2003; KRIEGER, 2006). Muitos desses nomes ficam restritos aos estudiosos da área (jargões), outros gradualmente tornam-se parte do vocabulário comum (PONTES, 1997; ISO 704, 2009; AHMAD, 2000).

[22] Termo utilizado para designar materiais artificiais que possuem propriedades não encontradas na natureza através da alteração da sua micro e macroestrutura ou da formação de um compósito.

A criação de termos científicos pode ser distinguida pelo seu caráter de neologismo ao nomear novos conceitos. Neologismo é um fenômeno linguístico que consiste na criação de uma palavra ou expressão nova (neologismo lexical), ou na atribuição de um novo sentido a uma palavra já existente (neologismo conceitual), que ainda não foi totalmente aceita na linguagem corrente (AHMAD, 2000; JAMET; TERRY, 2018).

Tendo isso em vista, o processo de constituir uma nomenclatura pode seguir alguns padrões. Compreende-se que termos científicos podem surgir: (i) completamente novos a partir de processos de derivação, composição, redução ou apropriação. Ou ainda, termos podem surgir (ii) a partir de palavras já existentes por processos de ressignificação, substantivação ou epônimos. Na sequência aprofunda-se na explicação.

Na derivação, a presença de prefixo ou sufixos de línguas clássicas como o latim, grego e árabe já foram mais comuns no processo de terminologia científica (HIRST, 2003), o que findava por facilitar muitas vezes a compreensão de um novo termo, pela familiaridade com os afixos (AHMAD, 2000). Um exemplo de derivação é a palavra 'robô' usada pela primeira vez para denotar um humanóide fictício em uma peça tcheca de 1920, R.U.R. (*Rossumovi Univerzální Roboti - Rossum's Universal Robots*) por Karel Čapek, e deriva de uma raiz eslava '*robot-'*, com significados associados ao trabalho (KURFESS, 2005). O processo de composição implica na junção ou fusão de duas, ou mais palavras, como em 'cyborg' = '***cyb****ernetic*' + '***org****anism'* (ISO 704, 2009, p.32). Além disso, reduções, siglas ou acrônimos[23], são bastante usados na ciência e podem vir a se comportar como novas palavras em nosso vocabulário, como se pode ver com o acrônimo 'radar', que abrevia '***Ra****dio* ***D****etecting* ***a****nd* ***R****anging*' (TIMBANE, 2014). A apropriação ocorre quando termos de outro idioma são assimilados para se referir a algo que não havia um termo similar, para exemplos recomenda-se o trabalho de Hoffer (2005) que apresenta historicamente diversas apropriações de variados idiomas pelo inglês.

Para os termos que surgem a partir de mudança semântica, ou seja, quando atribuí-se novo sentido a uma palavra já existente, confere-se um novo significado específico dentro de outro domínio. Como exemplo, apresenta-se a palavra 'circuito', que de modo geral significa "linha que limita inteiramente uma superfície; contorno, perímetro" (segundo dicionário

[23] A distinção entre sigla e acrônimo se dá pelo fato do acrônimo poder ser lido como uma nova palavra, e não necessariamente letra a letra, como na sigla. Embora ambos sejam formados a partir das letras iniciais de outras palavras (ISO 704:2009, p.43).

Michaelis[24]). No domínio da eletrônica, entretanto, um circuito adquire um significado particular, sendo um arranjo de dispositivos ou meios através dos quais a corrente elétrica pode fluir (ISO 704, 2009, p. 34). Ou ainda, essa mudança semântica pode estar relacionada à prática de se gerar acrônimos que retratam palavras comuns, como em SHRIMP e SQUID, que no vocabulário comum traduz-se por 'camarão' e 'lula', mas, respectivamente, correspondem à ***S**ensitive **H**igh-**R**esolution **I**on **M**icro**P**robe* e ***S**uperconducting **Q**uantum **I**nterference **D**evice,* equipamentos científicos de medição (SOVIĆ; BERTOŠA, 2009; TAY, 2020).

Não apenas, ideias podem ser nomeadas a partir de nomes de pessoas. Essa ocorrência é conhecida por epônimo, designando quando alguma coisa ou algum lugar, tem seu nome derivado do nome de uma pessoa ou personagem (KOSHLAKOV *et al.*, 2019). Na ciência é comum que descobertas ou inovações, recebam o nome do pesquisador ou de uma figura influente em seu avanço, fato conhecido por Lei de Stigler (STIGLER, 1980). Ainda que, no meio acadêmico, seja desencorajado a criação de novos epônimos, muitos ainda surgem ou persistem (HIRST, 2003). Isto pode ser observado com os termos 'Fordismo', sendo um processo nomeado derivado do nome próprio Henry Ford, assim como, o material 'baquelite' nomeado a partir de seu inventor Leo Baekeland. Ou ainda, o caso do cometa Halley que leva o nome de Edmond Halley, não porque o descobriu, mas porque foi o primeiro a prever seu comportamento.

No caso de descobertas na natureza, em zoologia ou botânica, assim como para novos elementos químicos ou descobertas em física e cosmologia, existem padrões e órgãos que regulam os processos de nomenclatura. Entre eles, citam-se o Código Internacional de Nomenclatura Zoológica (ICZN[25]) baseada na nomenclatura binomial de Lineu, a IUPAC[26] (*International Union of Pure and Applied Chemistry*) e a IUPAP[27] *(International Union of Pure and Applied Physics*). Embora a ISO 704, vise uma padronização para a comunicação do conhecimento, para o desenvolvimento tecnológico não há um padrão ou regulamentação vigente, tal que, novas palavras e termos para designar avanços tecnológicos podem surgir por diferentes trajetórias. O Quadro 8 traz uma síntese desses meios de formação de termos científicos identificados na literatura.

[24] https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/circuito/
[25] ICZN - https://www.iczn.org/the-code/the-code-online/?article=5
[26] IUPAC - https://iupac.org/what-we-do/nomenclature/brief-guides/
[27] IUPAP - https://iupap.org/who-we-are/internal-organization/commissions/c2-commission-on-symbols-units-nomenclature-atomic-masses-and-fundamental-constants/

Quadro 8 – Compilado de meios pelos quais uma nomenclatura pode surgir.

Origem do termo	Mecanismo	Descrição	Exemplo
Criação de novos termos	Derivação	Uso de sufixos e prefixos	'tigmotropismo'= do grego '*thixis*', que significa 'toque', e '*tropos*', que significa 'virar' ou 'direção'. Descreve especificamente o ato de uma planta ao entrar em contato com o objeto sólido e cresce ao seu redor (e.g., gavinhas do maracujá).
	Composição	Uso de justaposição ou aglutinação.	palavra '*smog*' = '*smoke*' + '*fog*' formada pela fusão de outras palavras, para representar o nevoeiro causado pela poluição.
	Redução	Uso de siglas ou acrônimos.	palavra 'laser', que abrevia '***L**ight **A**mplification by **S**timulated **E**mission of **R**adiation*'.
	Apropriação Transidiomática	Uso de termo ou expressão estrangeira que não exista equivalente.	palavra álgebra do árabe romanizado '*al-jabr*'.
Novos termos a partir de uma palavra já existente	Ressignificação	Uso da apropriação transdisciplinar ou acrônimos.	'vírus', da medicina, agente infeccioso. Para computação *malware*, um *software* malicioso. Ou ainda, o acrônimo SPIDER = *Spectral Phase Interferometry for Direct Electric-field Reconstruction*, também refere-se ao aracnídeo.
	Substantivação	Mudança na categoria sintática.	usando as terminações "*-ion/tion*", "*-ment*", "*-ity/–ty*", "*-ness*" e outros."(em inglês), verbos, adjetivos e advérbios são substantivados.
	Epônimo	Diz-se de, aquele ou aquilo, que dá o seu nome a outra coisa.	'motor Diesel', ou motor de ignição por compressão, tem seu nome atribuído ao engenheiro alemão Rudolf Diesel.

Fonte: Elaborado pela autora

Uma ressalva a ser feita nesta seção é que foram considerados processos de formação de palavras tomando a língua inglesa como referência. Isso se deu em razão do idioma inglês ser considerado como principal idioma científico (sendo também o idioma do conjunto de dados implementado no modelo proposto). No entanto, esses padrões de formação são particulares de cada idioma, como aponta Biderman (2001):

> "Assim os termos técnico-científicos são gerados com base na lógica da língua em questão, segundo os padrões lexicais nela existentes. Executam-se os empréstimos linguísticos, muito frequentes no mundo contemporâneo, sobretudo anglicismos, que se vêm propagando por todas as línguas, em virtude do papel hegemônico exercido pelos Estados Unidos na contemporaneidade. De fato, o inglês tornou-se a língua universal da ciência e da tecnologia" (BIDERMAN, 2001, p.15).

2.5 TRABALHOS RELACIONADOS

Essa seção tem o objetivo de evidenciar a contribuição de alguns estudos, mencionados previamente no referencial teórico, para fundamentação e elaboração do modelo proposto pela pesquisa relatada nesta tese. Compõem este conjunto de trabalhos relacionados, estudos teóricos e empíricos explicitamente associados ao tema de sinais fracos, além dos estudos sobre antecipação tecnológica e fronteira científica. Em virtude da escassez de estudos sobre sinais fracos para tecnologias emergentes buscou-se por informações em campos correlatos, nos quais, apesar de muitas vezes, não utilizarem o termo 'sinal fraco' demonstram interesse no objetivo de antecipar informações para tecnologia ou planejamento estratégico. Convém frisar

que nesta seção são considerados apenas aqueles estudos que visam a obtenção de informações por métodos quantitativos envolvendo técnicas de mineração de texto.

De início menciona-se o trabalho do JRC, "*Weak signals in science and technology*" (EULAERTS *et al*., 2019; 2021; 2022), sendo significativo por afirmar a relevância atual do tema de pesquisa, sendo um centro renomado de serviço de ciência e conhecimento da Comissão Europeia com o compromisso de apoiar as políticas da União Europeia visando impactar positivamente a sociedade. No mesmo sentido, menciona-se a publicação mais atual do KITSI - *Korea Institute of Science and Technology Information* (Yang *et al.,* 2022), na qual reconhecem a necessidade de abordagens para antecipação da informação tecnológica, ou seja, da busca por informações que não apresentem um histórico, noção que salientam ter adquirido a partir da publicação do JRC. Além disso, destaca-se a publicação de Magruk (2021) fortalecendo a relevância da busca por sinais fracos para tecnologia, explicitando a conexão entre o cenário de revolução industrial (indústria 4.0) elevando a incerteza relacionando com a teoria de sinais fracos de Ansoff. Assim como o estudo recente de Zhan e Du (2023) explicitando a relação entre sinais fracos e tecnologias emergentes, bem como o empenho na implementação de técnicas atualizadas de NLP.

Conceitualmente, sob o tema de sinais fracos, enaltece-se as publicações de Hiltunen (2008), Rossel (2009), Amanatidou *et al.* (2012) e Kim e Lee (2017), como fundamentais para a lógica embutida no modelo proposto. A obra de Hiltunen (2008) destaca-se por sua ampla adoção em estudos de sinal fraco, sistematizando o conceito de signo futuro (*future sign*) a partir do conceito de sinal fraco de Ansoff e da semiótica de Peirce. Assim, a autora define o signo futuro a partir de três dimensões: sinal, questão e interpretação (*signal, issue e interpretation*) (HILTUNEN, 2008). Colaborando para o entendimento dos aspectos que definem os sinais fracos, Rossel (2009) discorre sobre as perspectivas Ansoffianas e Construtivistas de entendimento de um sinal fraco. Essa construção de Hiltunen e a perspectiva Ansoffiana são essenciais para a proposta de identificação de sinais fracos desenvolvida nessa tese baseada em modelo quantitativo de descoberta de conhecimento em texto apresentada adiante no documento na seção 4.1.

Já os trabalhos de Amanatidou *et al*. (2012) e Kim e Lee (2017) são importantes por endossarem a discussão sobre a característica de antecipação (*earliness*) e novidade (*novel/novelty*) do sinal fraco. Para Amanatidou *et al.* (2012) um critério essencial para que os sinais sejam identificados como sinais fracos é que eles sejam desconhecidos (*unknown*). O que

sob o aspecto de mineração de texto, argumenta que as técnicas de mineração disponíveis na época não eram capazes de identificar "verdadeiramente" sinais fracos. Uma vez que as técnicas vigentes eram baseadas em rotinas estatísticas, o que implica em identificar partes do texto frequentemente mencionadas, implicando em selecionar uma informação relativamente conhecida. Kim e Lee (2017) contribuem com o conceito de "*novel future signal*", pontuando que sinais fracos sob aspecto antecipação podem não ser novos, mas sinais fracos sob o aspecto de novidade sempre serão sinais fracos.

A fim de facilitar esse entendimento, propõem-se a compreensão a partir do raciocínio de que um sinal fraco de novidade (*novel signal*) ocorre antes de um sinal fraco de antecipação (*earliness signal*), por exemplo, recentemente o estudo de "*weak signals*" poderia ter sido considerado um sinal fraco de antecipação de tema de pesquisa. Contudo, ele não era um sinal de novidade, pois seu conceito tem como marco teórico 1975. Dessa forma, nesse exemplo hipotético, "*weak signals*" seria um sinal fraco pelo aspecto de antecedência da informação de uma tendência emergente de pesquisa, mas não um sinal fraco de novidade como um novo tema de pesquisa. Assim, embora um sinal fraco seja sempre a respeito de uma nova informação, nem sempre será sobre uma novidade, mas uma novidade (desconhecido, *unknow*) sempre será um sinal fraco, ou um sinal fraco "verdadeiro" sob a perspectiva de Amanatidou *et al.* (2012).

A compreensão desse aspecto dos sinais fracos, em que ambos tratam de antecipar informações, mas uns antecedem outros, foi fundamental para considerar a demanda de antecedência "o quanto antes" colocada pela literatura de estudos antecipativos para tecnologia. E, assim, direcionar a pesquisa pela busca de uma solução capaz de identificar o fragmento de informação novo, desconhecido (sem precedente).

Dentre outros estudos, o trabalho de Ranaei *et al*. (2020), apesar de não usar o termo sinais fracos, é bastante explicito pontuando o desejo de identificar tecnologias quanto antes possível. Os autores enfatizam como as abordagens de modo geral estão presas a uma avaliação *ex-post*, consequentemente, enaltecem a falta de estudos providenciando métodos para identificação de novidades tecnológicas. Nesse sentido, entende-se que se a identificação de sinais fracos tem como objetivo antecipar tecnologias emergentes "o quanto antes", ele precisa ser um sinal fraco de novidade.

Considerando as abordagens empíricas visitadas na revisão da literatura, compreende-se que embora o termo sinal fraco seja empregado pelas publicações, em grande parte, são

dedicadas a buscar por novos termos de alta frequência (*emerging trend*) e poucos buscam por novos termos desconhecidos (*novelty*). Ou sequer esse tipo de discussão acerca da diferenciação entre abordagens de detecção de sinal fraco, para novidade ou antecipação de tendência, são explicitadas. A seguir detalham-se algumas abordagens julgadas relevantes.

O relatório JRC (EULAERTS *et al.*, 2019) fundamenta sua pesquisa em um sistema de mineração de texto, desenvolvido por eles, para acompanhar a evolução de tecnologias estabelecidas e emergentes chamado *Tools for Innovation Monitoring*® (TIM). No entanto, sua investigação baseia-se em um dicionário de termos obtidos pela frequência dos termos mais relevantes na base de publicações científicas *Scopus* e também no volume histórico de documentos que mencionam o termo, TF-IDF das palavras-chave. Nesse sentido, compreende-se que a pesquisa se concentra em identificar sinais fracos de antecipação de tendência, apontando que há espaço para aprimoramento na busca por sinais fracos de novidade.

O relatório KITSI (YANG *et al*., 2022), inspira-se na proposta da pesquisa do JRC, também extraem palavras da base de publicações científicas *Scopus* em um intervalo de 5 anos (ressaltam a escolha da base *Scopus* especialmente por conter anais de conferências, se mostrando adequado para detecção de sinais fracos, que precisam analisar as informações mais recentes). Implementam alguns parâmetros de frequência para monitorar e classificar as palavras que apresentam um movimento de destaque ao longo do tempo, denominando-as como "*popping keywords*". Em seguida, utilizam o modelo de *word embedding FastText*® (BOJANOWSKI *et al*., 2016) e medem a similaridade contextual entre as "*popping keywords*". Apresentam um método de agrupamento automático desenvolvido por eles definindo que o grupo de "*popping keywords*" derivado desse processo é definido como um sinal fraco. Eles oferecem como resultado a palavra-chave identificada como sinal fraco e as últimas 3 (três) publicações que apresentam essa palavra como palavra-chave para auxiliar na análise contextual do sinal fraco.

Do ponto de vista de investigar publicações científicas e/ou tecnológicas Ranaei *et al*. (2020) apostam na abordagem de minerar palavras/termos e se comprometem em investigar três abordagens a fim de investigar a característica emergente de uma área tecnológica visando antecipar o surgimento de novidades tecnológicas. Os métodos analisados foram: contagem de frequência de termos (TD-IDF), um indicador bibliométrico para emersão (*EScore*) e o modelo de tópicos *Latent Dirichlet allocation* (LDA). Uma crítica ao artigo reside no fato de desconsiderarem um termo quando ele não ocorre por dois anos consecutivos nos documentos.

Julga-se importante não desconsiderar os termos, pois em se tratando de novidades pode ser que demorem para ocorrerem outra vez, como no caso de publicações "*sleeping beauty*" (RAAN, 2004).

Um segundo ponto de crítica incide sobre os autores afirmarem que julgam satisfatória a implementação de métodos simples, como contagem de frequência, para detectar mudanças nas terminologias. Concorda-se que métodos de frequência são sensíveis e capazes de rastrear mudanças nas terminologias, no entanto, o questionamento é com qual antecedência isso poderia ser realizado? Quanto tempo seria necessário para que uma nova tecnologia possa ser detectada com base em frequência de palavras? Considera-se que, embora métodos baseados em frequência possam detectar o surgimento de novos termos, tendo em vista o cenário de aceleração tecnológica é importante atentar-se "quando" essa informação seria detectável. Dessa forma, evidencia-se a relevância de investigação de meios para fornecer "o quanto antes" esse tipo de informação sobre novidade tecnológica, como eles mesmos pontuam. Por fim, compreende-se que o estudo se concentra na identificação de novos termos de alta frequência, uma antecipação de tendência (*emerging trend*) e não de novidade, novos termos raros (*novelty*) como indicam ser a lacuna.

Ainda sobre a questão da novidade, partindo da relação entre ciência e tecnologia, a fim de atender essa demanda pela antecipação de informação, alguns estudos recorrem à investigação da fronteira da ciência. Shibayama, Yin e Matsumoto (2021) se dedicam a busca de novidade na ciência, no entanto, a metodologia proposta é presa a um vocabulário de novas palavras pré-fornecido, não sendo capaz de identificar uma nova palavra nem de atualizar seu *embedding*, apenas faz uso da biblioteca ScispaCy[28] para buscar a fronteira científica. Todavia, apresenta um esforço importante em implementar técnicas atualizadas de mineração de texto.

A complementar, muitos estudos que investigam ciência e tecnologia ao mencionarem que buscam por tecnologias emergentes (buscar uma informação) na realidade se comprometem em buscar emergência tecnológica (lacunas, ausência da informação). Embora a identificação de uma 'ausência de informação' seja uma informação nova e importante, podendo ser entendida como um sinal fraco para o desenvolvimento tecnológico (emergência tecnológica), configura um problema diferente de busca pela informação de novidade, um sinal fraco de novidade para tecnologia emergente. Essa discussão é aprofundada na seção 4.2.2.

[28] Baseada no modelo spaCy para representações vetoriais de vocabulários especializados em textos biomédicos.

Das pesquisas dedicadas a apresentarem abordagens para detecção de sinais fracos, Kim e Lee (2017) em seu estudo para sinais fracos de novidade para tecnologia propõem o uso de um modelo de detecção de *outliers* - LOF (*Local outlier factor*). Ele baseia-se em uma matriz palavra-chave/documentos e nos k-vizinhos mais próximos e como fonte de dados utiliza *websites* voltados para tecnologia e patentes. No entanto, não exploram em sua proposição a noção de temporalidade e desenvolvimento para identificação dos sinais. Isto fornece margem para a investigação sob esse aspecto de novidade a partir de publicações científicas e patentes, além de aprimorar as técnicas de mineração de texto.

Thorleuchter, Scheja e Poel (2014) se destacam por iniciarem a discussão da consideração semântica e temporal nos modelos de mineração de texto para identificação de sinais fracos. Por outro lado, fazem uso da análise de tópicos (LSI), uma abordagem que pode não ser eficiente quando se objetiva identificar um sinal de novidade com o máximo de antecedência, pois a análise de tópicos semântica necessita de tempo para conseguir identificar e agrupar assuntos semelhantes explicitados por expressões diferentes. Além disso, garante apenas que a informação seja nova, mas não, necessariamente, um sinal que expresse uma novidade.

Griol-Barres *et al.* (2020) também consideram a questão temporal em seu modelo apresentando um fluxo evolutivo para monitoramento do sinal fraco. O encadeamento proposto parte da mineração de publicações científicas, seguem com as reportagens jornalísticas e encerram com notícias divulgadas em rede social. Considera-se um ponto forte o aspecto temporal para acompanhamento evolutivo no processo de identificação dos sinais fracos. No entanto, Griol-Barres *et al.* (2020) baseiam sua etapa de mineração na frequência das palavras (DoD-DoV), e mais uma vez, reconhece-se como detecção de sinais fracos para antecipação de tendências e não de novidade.

Sob essa questão de monitoramento temporal, ressaltam-se os trabalhos de Dernis, Squicciarini e Pinho (2016) e Katsura e Ono (2019) ao proporem o conceito de *burst* para investigar o desenvolvimento científico e tecnológico e a fronteira científica, respectivamente. O conceito de *burst* é interessante por contribuir com o aspecto de antecedência, indicando atividades com aumento intenso em um determinado período de tempo em comparação com os níveis anteriores. Dernis, Squicciarini e Pinho (2016), ao implementam o conceito de *burst* a partir de termos nas bases de documentos de publicação científica e patentes, consideram apenas os títulos das publicações científicas da base *Scopus* e tratam cada palavra como um

token isolado, ou seja, não consideram a semântica do texto. Para as patentes a análise *burst* se deu sob os códigos de classificação IPC (*International Patent Classification*).

Por sua vez, Katsura e Ono (2019) investigaram o avanço científico com uma análise *burst* em uma rede dinâmica de palavras e nomearam sua abordagem de *TrendNets*®. A proposição baseou-se na frequência de coocorrência de palavras entre dois períodos de tempo sucessivos. Como mencionado anteriormente, embora, sejam capazes de explicitar informações novas, não se tratam de informações sobre novidades. De qualquer forma, contribuíram para consideração de temporalidade sugerindo a análise *burst* para monitoramento evolutivo das informações.

Outro achado interessante do estudo de Dernis, Squicciarini e Pinho (2016) foi a relação entre ciência e tecnologia ao observarem que para alguns campos tecnológicos, foi identificado primeiro um comportamento *burst* nas patentes seguido de um comportamento *burst* nas publicações científicas. E, não o contrário, o que seria mais intuitivo, considerando o modelo de inovação linear, onde o conhecimento flui da ciência básica (publicações científicas) para a aplicação técnica (base de patentes). Essa observação em conjunto com a noção de publicações do tipo "*sleeping beauty*", na qual Rann (2017) afirma que certas publicações são primeiramente verificadas em citações nos documentos de patentes do que identificadas nas redes de citação de publicações científicas. Assim, contribuem para o entendimento que apenas investigar o avanço de fronteira científica como fazem, Shibayama, Yin e Matsumoto (2021), não é suficiente para antecipar informações de novidade para o domínio tecnológico, sendo necessário investigar e relacionar as informações presentes tanto na base científica quanto na base de conhecimento tecnológico.

Por fim, comenta-se sobre os estudos mais recentes de identificação de sinais fracos sob o aspeto de técnicas de mineração de texto. Ebadi, Auger e Gauthier (2022), Yang *et al*., (2022) e Zhan e Du (2023), são estudos que implementam técnicas de NLP buscando explorar aspectos linguísticos, recorrendo aos modelos de *embedding* BERT e FastText. No entanto, o uso dessas técnicas recai sob a tarefa de expandir a compreensão do sinal fraco (*topic* ou *keyword co-occurrence perspective*), mas não para identificá-lo. A fundamentação para identificar os sinais fracos ainda são abordagens estatísticas com base em frequência, dando margem para explorar a potencialidade semântica dos recursos atuais de NLP.

Zhan e Du (2023) até o momento é o trabalho mais recente de identificação de sinais fracos, e particularmente, pontua-se que o trabalho também se trata da identificação de sinais fracos para tecnologias emergentes, validando o percurso de lacuna teórica desenvolvido nesta tese. Sobretudo, os autores trazem uma crítica relevante da falta de contextualização dos sinais fracos ofertados pela abordagem estatística (DoV; DoD) baseada em palavras-chave. Nesse sentido implementam a mesma abordagem de mapas de palavras-chaves, mas ao identificar as palavras-chaves consideradas sinais fracos obtém seus *embeddings* pelo modelo BERT e realizam um cálculo de similaridade por cossenos para identificar outras palavras-chaves associadas ao seu contexto. Então, assim como na abordagem do KITSI (YANG *et al*., 2022), demonstram preocupação em contextualizar os sinais fracos encontrados para facilitar o processo de interpretação e compreensão do significado desses sinais. Entretanto, ainda apresentam uma abordagem baseada da abordagem de mapas de frequência (DoV; DoD) apresentada por Yoon (2012) para detectar termos tidos como sinais fracos.

Como conclusão, recupera-se a ponderação de Amanatidou *et al.* (2012), que já mencionava a necessidade de técnicas mais sofisticadas para capazes de encontrar sinais fracos de "verdade", sob o aspecto de identificar o desconhecido. O avanço dos métodos baseados na mineração de texto abriram novas oportunidades no processo de detecção de sinais fracos, tecnologias emergentes e fronteira científica. Compreende-se, portanto, nesta tese o objetivo da mineração de texto como sendo detectar informações previamente desconhecidas como solução para o problema de detecção de potenciais tecnologias emergentes. Sob essa perspectiva, visou-se explorar a potencialidade das técnicas atualizadas para mineração de texto baseadas em técnicas NLP e ML, permitindo que a semântica os textos fossem consideradas no modelo proposto, além de ser considerada também a questão temporal e dinâmica por meio da análise *burst,* explorando também, mais de um tipo de fonte de dados para a detecção de sinais fracos.

2.6 SÍNTESE DO CAPÍTULO

A fim de sintetizar a lógica no encadeamento das teorias apresentadas neste capítulo, é explicitada a razão pela qual as abordagens utilizadas conduzem a uma nova questão teórica. Neste sentido, são três eixos teóricos principais os quais sustentam esta tese: (i) Sinais Fracos, (ii) Estudos Antecipativos para Tecnologia e (iii) Descoberta de Conhecimento em Texto. O primeiro faz parte do referencial de Estudos Futuros para Planejamento Estratégico, conjunto de perspectivas teóricas que se ocupam da antecipação de informação. O segundo, como

recorte, aborda o contexto tecnológico para a aplicação da detecção de sinais fracos, visto sua importância para o desenvolvimento econômico-social e o atual momento de mudanças tecnológicas acentuadas (Indústria 4.0). Por fim, o terceiro diz respeito ao arcabouço de técnicas computacionais para amparar o processo de identificação dos sinais fracos.

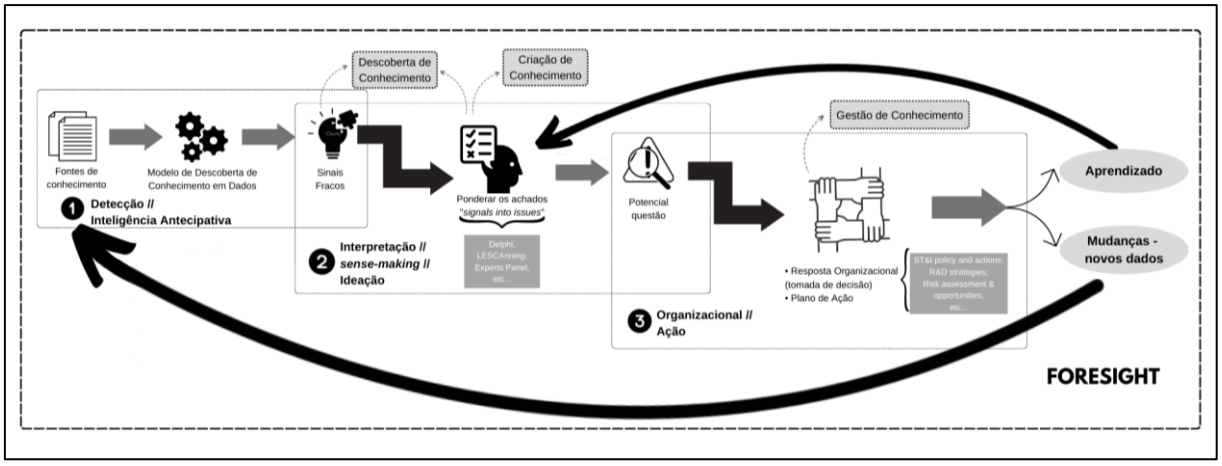
Sinais fracos para antecipação de informação sobre novidade tecnológica é guia principal da investigação desta tese. Dentro desse âmbito foca-se na detecção de sinais fracos para tecnologias emergentes. Esta tese defende que os sinais fracos de novidade para tecnologias emergentes podem ser identificados por meio de modelos de descoberta de conhecimento em texto, resultante da relação entretecnologia, publicações científicas e pedidos de patentes, sendo onde o conhecimento se situa.

Sobretudo, para a etapa de detecção de sinais fracos, considerando o volume de dados e a interferência do viés cognitivo no uso de especialistas, se faz necessário recorrer a modelos de descoberta de conhecimento em texto. Argumenta-se, então, que à medida que novos conhecimentos são gerados e publicados, esses fragmentos de informação podem ser detectados, visto que o conhecimento ocorre dentro de um certo domínio, permitindo identificar e monitorar como se relacionam e evoluem. Ademais, o cenário de aceleração tecnológica marcado pela quarta revolução industrial urge pela antecipação de informações, logo destaca-se a relevância do estudo tomando como premissa que sinais fracos de novidade antecedem sinais fracos de antecipação de tendência tecnológica.

Por fim, investiga-se como se dá a formação de novas palavras tomando como premissa a compreensão de palavras como o menor nível de abstração de uma informação. Visando satisfazer a lacuna de pesquisa de antecipação de informação o quanto antes e explorar a semântica presente na parte não estruturada dos documentos textuais.

A Figura 7 ilustra como o modelo proposto (etapa 1 de Detecção) contribui para um *framework* maior de *Foresight*. Convém destacar que a Figura 7 não se trata de um modelo a ser perseguido ou validado. O intuito deste esquema é localizar onde o modelo proposto se encontra dentre o conhecimento sobre *Foresight*. A tarefa de detecção de sinais fracos para tecnologias emergentes contribui para a função de Inteligência Antecipativa em um processo de *Foresight*. Demonstrando, então, a utilidade no contexto da área fim a qual o artefato (modelo) se propõe a colaborar.

Figura 7 – Esquema representativo do processo de pesquisas antecipativas (*Foresight*).



Fonte: Elaborado pela autora

3 MÉTODO DE PESQUISA

Este capítulo abarca o processo de pesquisa que originou esta tese. A seção 3.1 expõe a visão de mundo e classificação da pesquisa. Em seguida, a seção 3.2 apresenta a *Design Science Research Methodology* (DSRM), abordagem que direciona os procedimentos metodológicos adotados e descritos na seção 3.3.

3.1 CLASSIFICAÇÃO DA PESQUISA

A visão de mundo do pesquisador é a base para a pesquisa científica. Ela reflete aquilo que se compreende e reconhece por conhecimento científico e a forma como o acessa, em outras palavras, suas bases ontológicas e epistemológicas. O paradigma de pesquisa é o que se entende por esse conjunto de crenças e acordos comuns compartilhados entre pesquisadores sobre como os problemas devem ser compreendidos e tratados (KHUN, 1962). A metodologia é onde as suposições sobre a natureza da realidade e do conhecimento, teoria e prática sobre um determinado tópico, se conetam (CRESWELL, 2010). De modo sucinto, pode-se definir a pesquisa científica como uma investigação sistemática de um determinado assunto a fim de obter novas conclusões, um conhecimento novo e confiável sobre o objeto de pesquisa. Por conseguinte, é importante ter clareza sobre o paradigma no qual a pesquisa se enquadra, pois isso orienta e traz implicações para as decisões tomadas no processo de pesquisa (CRESWELL, 2010).

À vista disso, a presente pesquisa se classifica, quanto à sua natureza, como uma pesquisa aplicada, a qual, conforme Silva e Menezes (2001, p.20), "objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos". Mais que isso, trata-se de uma pesquisa tecnológica, isto é, fundamenta-se no conhecimento tecnológico intrínseco ao "estudo científico do artificial", difundida também como *Design Science*[29] (DS) (SIMON, 1996).

[29] O termo poderia ser traduzido por algo como "Ciência de Projeto", entretanto, optou-se pelo emprego do termo original "*Design Science*" em virtude da variedade de traduções que se pode atribuir ao termo "*Design*". Segundo o Manual de Oslo (OECD) (1997, p.23) *"A palavra design, na língua inglesa, pode ter diferentes interpretações, além da mais conhecida pelos brasileiros ligada a estilo, moda, layout do produto. As demais acepções dessa palavra aparecem neste Manual e são traduzidas pelos seus sentidos. Empregam-se assim, além da palavra 'design', as palavras 'concepção', 'desenho', 'delineamento' e 'formulação'.*" Assim, em função de seu sentido polissêmico não a traduzimos.

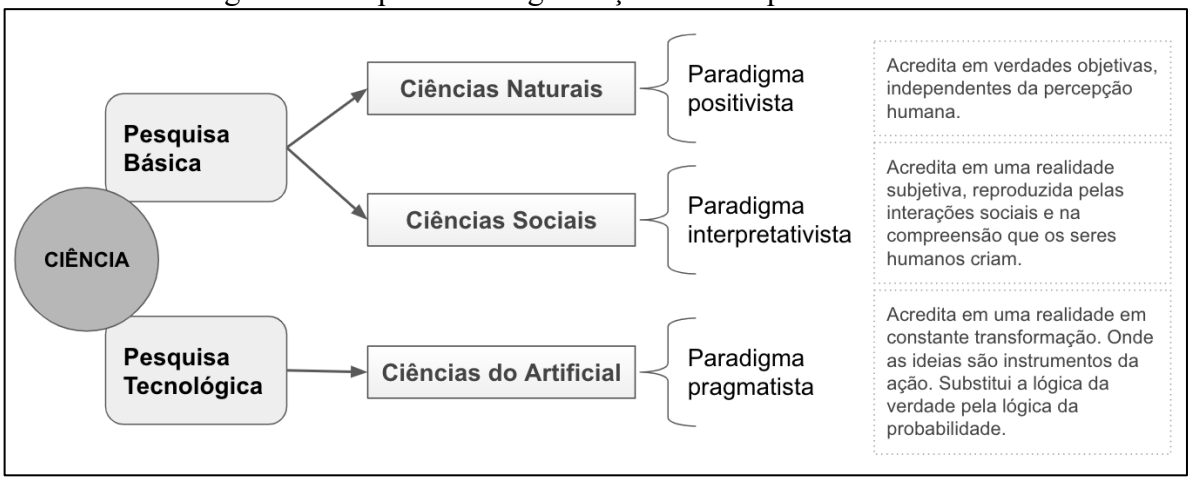
Cabe destacar que o entendimento da *Design Science* enquanto paradigma ou base filosófica ainda está em desenvolvimento e adoção, sendo relativamente recente em comparação à abordagem de pesquisa científica tradicional (PURAO, 2013). A *Design Science*, como ciência do artificial, tem a publicação de 1969 de Herbert Simon: "*The Sciences of Artificial*", como principal marco teórico. O autor argumenta em prol do conhecimento tecnológico como cientificamente válido e sobre a necessidade de uma ciência que se dedique a propor como construir artefatos, gerando conhecimentos com validade científica. Cupani (2016) respalda essa visão de conhecimento tecnológico ao ressaltar que, por definição, a tecnologia é um fenômeno que pertence ao âmbito do conhecimento; a '*techne'*, do grego, não é um mero fazer, mas um saber-fazer. Nesse sentido, a *Design Science* caracteriza o processo científico de construção do conhecimento do saber-fazer.

Simon (1996) define como artefato tudo o que não é natural, ou seja, algo construído pelo ser humano. A complementar, para Cupani (2016, p.171), "o artificial constitui um sistema adaptado ao ambiente em função de um determinado propósito humano, um objeto (artefato) com propriedades desejadas, idealizado e fabricado conforme um projeto (*design*)". Relevante notar que essa interpretação vai além da associação usual entre artefato como produto (artefato físico), significando inclusive o engendro de conhecimento aplicável e útil para a solução de problemas (artefato intelectual). Salienta-se, todavia, que a concepção de artefatos que realizem objetivos, por si só, não se caracteriza como uma pesquisa científica, conforme frisado por Wazlawick (2014), o que o autor declara como um caso de '*design* puro'. À vista disso, como concluem Hevner *et al*. (2004), a *Design Science* tem como indispensável, além do projeto de um artefato, o avanço do conhecimento. Assim sendo, a DS não se preocupa com a ação em si mesma, mas com o conhecimento utilizado para projetar e os novos conhecimentos gerados pelo processo e pelas soluções. Por essas razões é vista como uma base epistemológica.

Diante desse cenário, com intuito de elucidar essa visão de mundo, situa-se a DS em relação à pesquisa científica tradicional. De maneira breve, explica-se: enquanto as ciências naturais têm como foco de análise fenômenos da natureza, buscando entender 'como' e 'porquê' as coisas são como são, sob um paradigma positivista. E, as ciências sociais prezam pelos fenômenos sociais e buscam refletir sobre o ser humano e suas ações, sob um paradigma interpretativista. Analogamente, a *Design Science* é orientada à solução de problemas e se dedica a projetar e produzir algo que ainda não existe. É a ciência do artificial, prescritiva, sob o paradigma pragmático, que atua modificando a situação vigente para alcançar melhores

resultados com foco na solução (CUPANI, 2016; SIMON, 1996). Um esquema dessas ideias é apresentado na Figura 8.

Figura 8 – Esquema de organização das Pesquisas Científicas.



Fonte: Elaborado pela autora

A DS apresenta, ainda, uma particularidade distinta às ciências tradicionais, ao recorrer à lógica abdutiva como forma de inferência. Nessa perspectiva, se por um lado, a pesquisa científica básica explora algo existente e faz uso dos raciocínios indutivos e dedutivos, por sua vez, a pesquisa tecnológica objetiva produzir algo novo fundamentado no raciocínio abdutivo (CUPANI, 2016). A indução e a dedução são as lógicas de raciocínio aplicadas nas ciências tradicionais, associadas, respectivamente, a explorar ou explicar, e a descrever ou predizer (SACCOL, 2009). Já a abdução, como resultado do trabalho de Charles S. Peirce, é a inferência a favor da melhor explicação; explora dados, identifica padrões, sugere hipóteses e permite novas descobertas em prol de criar explicações para um determinado fenômeno (PEIRCE, 1972). Como efeito, a abdução se diferencia pelo raciocínio de inferência que se projeta para o futuro, enquanto a dedução e a indução se referem ao passado, ao já conhecido, na medida em que se referem à experiência.

Como mencionado, a *Design Science* quanto paradigma de pesquisa, tem seu fundamento no pragmatismo. Isso significa que acredita que a realidade é constantemente ajustada, debatida, interpretada e, portanto, a melhor abordagem a ser utilizada é aquela que resolve o problema, sendo a questão de pesquisa o determinante mais importante nesse paradigma de pesquisa (MENEGHETTI, 2007). Nesse contexto, o pragmatismo é o processo de abdução, manter-se aberto às múltiplas soluções, associado à criação, ao engendramento de uma nova forma de fazer, gerando conhecimento. Em suma, a DS se concentra em projetar e prescrever. Nas palavras de Simon (1969, p.55, tradução nossa), “enquanto as ciências naturais

e sociais tentam entender a realidade, a DS tenta criar coisas que servem a propósitos humanos". A concluir, a *Design Science* tem a capacidade de unir a habilidade da ciência de entender "o que é", e a do *design* de entender "o que pode ser" (HEVNER; CHATTERJEE, 2010).

Por fim, considerando a orientação teórica da *Design Science,* visando a concepção de um artefato, que neste projeto de tese tem a Descoberta de Conhecimento em Texto (KDT) como base, implica em adotar uma abordagem quantitativa. A qual "considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las" (PRODANOV; FREITAS, 2013, p.69). Para tanto, recorre-se a sistematização da *Design Science Research Methodology* (DSRM), proposta por Peffers *et al.* (2007), como procedimento metodológico, apresentada na seção seguinte. O Quadro 9 expõe uma síntese da classificação da pesquisa quanto à sua natureza, abordagem, objetivo e procedimentos.

Quadro 9 – Síntese da classificação do projeto de pesquisa da tese.

Classificação da pesquisa quanto à:	
Natureza	Tecnológica
Abordagem	Quantitativa
Objetivo	Projetar e Prescrever
Procedimento	*Design Science Research Methodology*

Fonte: Elaborado pela autora

3.2 DESIGN SCIENCE RESEARCH METHODOLOGY

A condução da pesquisa segundo a DS, enquanto paradigma científico, é dada pelas abordagens de *Design Science Research* (DSR) fornecendo estrutura para sua realização. A abordagem DSR enquanto procedimento metodológico se despontou e obteve reconhecimento inicialmente nas pesquisas da área de Sistemas de Informação (SI) (HEVNER *et al.*, 2004; MARCH; SMITH, 1995). A DSR pode ser compreendida como um conjunto de diretrizes para conduzir um processo metodológico que permita a operacionalização das pesquisas em DS, alinhada à necessidade de procedimentos racionais e sistemáticos para a condução de projetos de novos artefatos tecnológicos (LACERDA *et al.*, 2013). Nesse sentido, a *Design Science* é a base epistemológica ao passo que a *Design Science Research* é a metodologia que operacionaliza a construção do conhecimento neste contexto (CHAKRABARTI, 2010). Desde

então, a implementação da DSR tem se expandido para outros domínios, até mesmo nas ciências sociais (LEE; LEE, 2019).

A DSR, enquanto metodologia que orienta o percurso a ser realizado na pesquisa, é competente, tanto na construção de artefatos, quanto na construção de conhecimento. Dresch, Lacerda e Antunes (2015) evidenciam a importância não apenas da solução específica desenvolvida pela DSR, mas do conhecimento envolvido, ao mencionar a contribuição do artefato para "classes de problemas", possibilitando assim, a generalização e o avanço do conhecimento.

Sob essa ótica, Wieringa (2009) explicita a contribuição científica da DSR, compreendendo a metodologia como uma estrutura lógica para resolução de problemas, que consiste em dois ciclos distintos, executados paralelamente de forma complementar, deliberando sobre dois tipos de problema: (i) o ciclo regulador, focado na concepção do artefato, nas atividades de engenharia, associado aos "problemas práticos", e que alteram o estado do mundo e geram conhecimento; e (ii) o ciclo do conhecimento, associado aos "problemas de conhecimento", que demandam uma mudança a respeito do conhecimento sobre o mundo, alterando o estado do conhecimento e aplicando-o no mundo. Esta inter-relação de problemas de *design* e problemas de pesquisa em um mesmo processo metodológico é o que caracteriza a pesquisa em *Design Science* voltada à produção de conhecimento científico por meio do desenvolvimento e uso de artefatos.

Como resultado da pesquisa conduzida pela DSR tem-se um artefato, que segundo descrevem March e Smith (1995), podem ser:

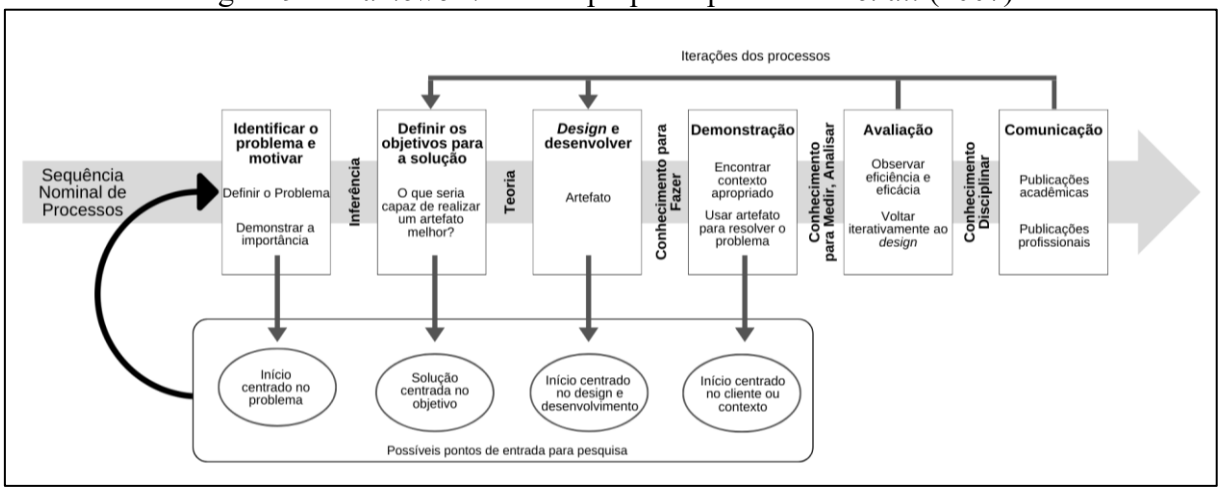
- Constructos: conceitos que formam o vocabulário de um domínio, definindo os termos usados para descrever e pensar sobre as tarefas.
- Modelos: conjunto de proposições que expressam relacionamentos entre constructos, descrevendo ou representando objetos do mundo real.
- Métodos: conjunto de passos usados para executar determinada tarefa.
- Instanciações: concretização de um artefato em seu ambiente, demonstrando a viabilidade e a eficácia dos modelos e métodos. Informam ainda como implementar ou utilizar determinado artefato e seus possíveis resultados.

O *framework* DSRM (*Design Science Research Methodology*) proposto por Peffers *et al*. (2007), originalmente voltado aos estudos em SI, é uma das metodologias seguidas mais

difundidas[30] para operacionalizar a pesquisa em DS. Seu desenvolvimento ocorre a partir de seis etapas procedurais: (1) identificação do problema e motivação; (2) definição dos objetivos para a solução; (3) *design* e desenvolvimento; (4) demonstração; (5) avaliação e; (6) comunicação. Cada etapa pode ser viabilizada pela implementação de diferentes métodos e técnicas.

O modelo de processos apresentado por Peffers *et al*. (2007) permite ainda, quatro possíveis alternativas de entrada para o percurso da pesquisa; motivadas e baseadas no problema, nos objetivos, na modelagem e desenvolvimento, ou, ainda, diretamente voltada para o cliente/contexto (Figura 9).

Figura 9 – *Framework* DSRM proposto por Peffers *et al.* (2007).



Fonte: Traduzido de Peffers *et al*. (2007)

Assim como para as demais abordagens de pesquisa científica, a DSR requer um conjunto de cuidados, tanto na construção como na avaliação do resultado da pesquisa. Esse rigor é fundamental para a confiabilidade do projeto do artefato (LACERDA *et al.*, 2013; PURAO *et al.*, 2008). Na DSR, seguindo seu fundamento pragmático, a avaliação não tem como objetivo mostrar "por que" ou "como" o artefato funciona, mas observar e medir o quão bem o artefato se suporta como uma solução para o problema (PEFFERS *et al.*, 2007). Faz parte desse processo averiguar tanto a validade científica, ou seja, o rigor na concepção e condução da pesquisa, como a validade funcional, se o artefato atende à solução do problema identificado.

A escolha do método de avaliação deve estar alinhada diretamente ao artefato em si e sua aplicabilidade (PEFFERS *et al*., 2007). Com isso em mente, mensura-se como o artefato

[30] Próximo de 10 mil citações registradas no Google Acadêmico 2022 <https://scholar.google.com.br/scholar?q=%22Design+Science+Research+Methodology>

atende à solução do problema comparando, por meio de métricas e técnicas de análises, os objetivos propostos para a solução com os resultados observados na utilização do artefato durante a etapa de avaliação, que pode ser em um ambiente experimental ou real, como pontuam Lacerda *et al*. (2013). Na visão de Hevner *et al*. (2004, p.86), os artefatos podem ser avaliados de forma experimental, analítica, observacional, teste ou descritiva, sendo que cada forma de avaliação conta com métodos e técnicas específicas (Quadro 10). Em poucas palavras, o processo de avaliação pode incluir qualquer evidência empírica adequada ou prova lógica.

Quadro 10 – Métodos de avaliação para a DSR.

Forma de Avaliação:	**Métodos propostos:**
Observacional	Estudo de Caso: Estudar o artefato existente, ou não, em profundidade no ambiente de negócios. Estudo de Campo: Monitorar o uso do artefato em projetos múltiplos. Esses estudos podem, inclusive, fornecer uma avaliação mais ampla do funcionamento dos artefatos configurando, dessa forma, um método misto de condução da pesquisa.
Analítico	Análise Estatística: Examinar a estrutura do artefato para qualidades estáticas. Análise da Arquitetura: Estudar o encaixe do artefato na arquitetura técnica do sistema técnico geral. Otimização: Demonstrar as propriedades ótimas inerentes ao artefato ou então demonstrar os limites de otimização no comportamento do artefato. Análise Dinâmica: Estudar o artefato durante o uso para avaliar suas qualidades dinâmicas (por exemplo, desempenho).
Experimental	Experimento Controlado: Estudar o artefato em um ambiente controlado para verificar suas qualidades (por exemplo, usabilidade). Simulação: Executar o artefato com dados artificiais.
Teste	Teste Funcional (*Black Box*): Executar as interfaces do artefato para descobrir possíveis falhas e identificar defeitos. Teste Estrutural (*White Box*): Realizar testes de cobertura de algumas métricas para implementação do artefato (por exemplo, caminhos para a execução).
Descritivo	Argumento Informado: Utilizar a informação das bases de conhecimento (por exemplo, das pesquisas relevantes) para construir um argumento convincente a respeito da utilidade do artefato. Cenários: Construir cenários detalhados em torno do artefato, para demonstrar sua utilidade.

Fonte: Elaborado a partir de Hevner *et al.* (2004, p.86)

3.3 PROCEDIMENTOS METODOLÓGICOS

Como resultado desta pesquisa de doutorado foi concebido, como artefato, um modelo de descoberta de conhecimento em texto para detecção de sinais fracos para tecnologias emergentes. Como processo metodológico, seguiram-se as etapas da DSRM proposta por Peffers *et al.* (2007). Considerando o fluxo apresentado na Figura 9, o ponto de entrada do modelo ocorreu centrado no problema, ou seja, a ideia da pesquisa foi consequência da constatação do problema por meio da análise dos artigos científicos em conjunto com as

recomendações para pesquisas futuras presentes na literatura visitada. Na continuidade, foram documentados os processos envolvidos em cada etapa da DSRM.

3.3.1 Identificar o Problema e Motivar

A primeira etapa se dedica a identificação e definição do problema e das justificativas que demonstram o valor da solução proposta. Em função disso, se faz necessário o conhecimento do estado da arte do problema, podendo-se recorrer a diferentes técnicas para esse propósito.

O ponto de entrada da pesquisa, como mencionado anteriormente, ocorreu nesta primeira etapa por constatação de lacuna teórica. Isso significa que o primeiro passo realizado foi uma busca estruturada exploratória na literatura. Buscou-se por conteúdos acerca do tema de identificação de tendências motivado pelo interesse no tema de antecipação de conhecimento e estudos prospectivos sobre o futuro. Nesse primeiro momento, identificou-se a demanda por aqueles fragmentos de informação mais incipientes, os sinais fracos. Bem como a importância desse tipo de pesquisa para o domínio tecnológico, tendo em vista o contexto recente de avanço tecnológico e Indústria 4.0.

Sob essa perspectiva, em um segundo momento, aprofundou-se na investigação sobre o conceito de sinal fraco. Uma segunda revisão da literatura foi realizada, tratando especificamente sobre os temas de "*weak signals*", "*technology foresight*" e "*emerging technology*", já considerando o recorte das tecnologias emergentes como contexto de interesse.

Cabe ressaltar que, o estudo de sinais fracos é um tema de interesse atual, em crescente desenvolvimento, como consequência, não apresenta um *corpus* de conhecimento claramente definido. Nesse sentido, apesar de se considerar um protocolo estruturado de busca, a técnica de revisão implementada foi a revisão narrativa. A título de exemplo, uma busca na base de publicações *Web of Science* (WoS) com os seguintes descritores: (*"weak signal" and "technolog*" and "foresight"*) retornou 9 artigos, sendo que dois deles podiam ser descartados por apresentarem uma vertente social para sinais fracos e não tecnológica. Em comparação, uma busca apenas com os descritores (*"weak signal" and "technolog*"*) na mesma base WoS, retornam 1074 publicações, entretanto, uma proporção massiva desses documentos recuperados podia ser descartada, pois, diziam respeito a outro tipo de sinal fraco, como ruído em sistemas.

Essa “interferência” na pesquisa ocorre em razão do termo sinal fraco adotado por Ansoff derivar da teoria da informação, ocasionando, muitas vezes, na associação do termo à estudos de eletrônica, captação e processamento de sinais (ondas) em engenharia ou neurociência. Além disso, mesmo com o intuito de referenciar a noção de sinal fraco proposta por Ansoff, existe uma pluralidade de expressões que remetem a essa ideia como visto no referencial teórico. Isso dificulta a coleta sistemática de publicações que relacionam o tema de sinais fracos e tecnologia sob a perspectiva de Ansoff e do planejamento estratégico antecipativo. Por essa razão, a busca pela literatura foi efetuada de forma narrativa e não sistemática.

Isto posto, embora a revisão se classifique como narrativa, tomou-se a busca estruturada de revisão integrativa como ponto de partida, assim, diferentes buscas foram efetuadas combinando dois conjuntos de descritores (Quadro 11). Agregando também estudos presentes nas referências desses trabalhos para compor o arcabouço de publicações utilizadas como fundamentação teórica para a pesquisa. Foram consultadas as bases *Web of Science ; Scopus®; ACM®; IEEE Xplore®; SpringLink® e ScienceDirect®*.

Quadro 11 – Termos usados para obter a bibliografia utilizada para fundamentar a pesquisa.

Principais descritores usados para buscar por estudos de sinais fracos:	*weak signal; early signal; future signal; wild card; emerging issue; novel signal; novelty detection; seeds of change; warning signal; early detection.*
Principais descritores usados para delimitar o recorte de interesse:	*foresight; technology foresight; technology; technologies; technological; emergent technolog*; emerging technolog*; technolog* emergence; anticipation; future-oriented.*

Fonte: Elaborado pela autora

Nesse momento, identificou-se a classe de problema: “detecção de sinais fracos” e delimitou-se o contexto de aplicação “para tecnologias emergentes”, culminando na definição do nicho de pesquisa: “Como detectar sinais fracos para tecnologias emergentes a partir de uma abordagem quantitativa?”. Pode-se notar que em decorrência da pesquisa bibliográfica houve uma alteração na motivação inicial atrelada a detecção de tendências, para a detecção de novidade a partir do conceito de sinais fracos, conforme apontado como lacuna pelas pesquisas. O levantamento da literatura possibilitou também reunir justificativas para a relevância desse problema e base teórica para a proposta da solução.

De forma sucinta, a pesquisa está posicionada no campo de Estudos Futuros e *Foresight.* Sob o recorte de antecipação de informação para o planejamento estratégico, podendo ser descrita também como inteligência antecipativa (*Anticipatory Intelligence*), com o problema de detectar sinais fracos para tecnologias. Em particular, possui o propósito de desenvolver uma abordagem quantitativa para detectar sinais fracos para tecnologias emergentes.

3.3.2 Definir os Objetivos para a Solução

Apoiada na etapa anterior, esta etapa concentra-se na definição dos objetivos da solução pretendida. A partir do estudo em profundidade da literatura reunida, ponderou-se sobre o que seria possível e viável, considerando as soluções já existentes, suas características e elementos, e sugestões de estudos futuros, para estabelecer o escopo da proposta de solução.

Primeiro, sinal fraco é um conceito que ainda não apresenta uma definição consolidada. Como efeito, diversos métodos e técnicas podem ser encontrados na literatura, qualitativos e quantitativos. Contudo, este trabalho foca nas abordagens quantitativas, considerando a capacidade dos modelos de descoberta de conhecimento em texto para obter fragmentos de informação. Sobretudo, evidencia-se a relevância da abordagem computacional quantitativa da mineração de texto para o tema, ao viabilizar a automação da detecção desses sinais, tendo em vista o volume de dados a serem analisados e, também, minimizar a influência de vieses cognitivos inerentes ao uso de especialistas para a tarefa.

Segundo, considerando as abordagens quantitativas, foi possível identificar aspectos a serem contemplados pelo artefato elaborado, não abordados até então. Destacam-se como objetivos para a solução as contribuições futuras mencionadas nos estudos visitados na revisão bibliográfica: considerar a semântica do texto, a questão da temporalidade e dinamicidade do modelo e valer-se de diferentes bases de dados para as análises (MÜHLROTH; GROTTKE, 2018; ROUSSEAU; CAMARA; KOTZINOS, 2021; ZHAO, TANG, HE, 2023).

Tendo isso em vista, a etapa de pesquisa teórica foi complementada com uma revisão da literatura sobre modelos, métodos e técnicas de descoberta de conhecimento em texto, mineração de texto, processamento de linguagem natural, aprendizado de máquina e análise de redes. Bem como, buscou-se embasamento teórico na temática de tecnologias emergentes, antecipação tecnológica, semiótica, formação de neologismos e terminologia científica, visando determinar como o artefato poderia dar suporte as metas estabelecidas.

Como conclusão, definiu-se como objetivo para a solução, a necessidade de um modelo quantitativo para detecção de sinais fracos, capaz de realizar a investigação semântica dos textos, que inclua a dinamicidade e temporalidade das informações em sua análise e utilize mais de uma base de conhecimento. Uma síntese das necessidades e proposições para solução do problema podem ser vistas no Quadro 12.

Quadro 12 – Síntese do escopo da solução proposta.

Necessidades	Proposições
Lidar com volume de dados e viés cognitivo de especialistas	Automação por abordagem computacional quantitativa de Descoberta de Conhecimento em Texto
Antecipação da informação "o quanto antes"	Encontrar novas palavras como *proxy* de tecnologias emergentes
Considerar a semântica dos textos	Técnicas de e*mbeddings* contextuais dinâmicos de palavras
Considerar temporalidade e dinamicidade do modelo	Análise de redes temporais permitem atualização e reaproveitamento do modelo
Uso de mais de uma fonte de conhecimento	Uso de bases científicas e tecnológicas para o conjunto de dados analisado pelo modelo

Fonte: Elaborado pela autora

3.3.3 Projetar e Desenvolver Artefato

Etapa central com a finalidade de criar o artefato. De modo geral, o projeto deve contemplar a designação da funcionalidade e arquitetura do artefato, incorporando o conhecimento teórico previamente assimilado. Podendo empregar diferentes técnicas julgadas necessárias para concretizar os objetivos definidos para o artefato.

Face às possibilidades de artefatos descritos por March e Smith (1995), como objetivo desta tese, foi estabelecida a proposição de um modelo como solução para o problema de detecção de sinais fracos para tecnologias emergentes. Visto que elaborar um modelo implica em conceber um encadeamento conceitual de processos que podem ser realizados por diferentes técnicas. Com base na literatura, estabeleceram-se as tarefas que compõem o artefato proposto.

Para elaborar o modelo, além das características levantadas na etapa anterior tomaram-se como características basilares, a estrutura de um modelo KDT (Figura 4), a premissa estabelecida por Ansoff de uma sequência de ocorrências a serem identificadas e acompanhadas a fim de detectar sinais fracos, como também, os três aspectos fundamentais dos sinais fracos:

novidade, incerteza (fragmento vago) e relevância estratégica (potencial de impacto), para engendrar a arquitetura do modelo e a capacidade de atender as necessidades identificadas.

O modelo, como proposta de solução, foi desenvolvido considerando sua adaptação à medida que as técnicas de processamento de linguagem natural e a capacidade de processamento computacional evoluem. Além disso, levou-se em conta a possibilidade de transposição do modelo para um cenário diferente do contexto tecnológico ao aplicá-lo à diferentes bases de conhecimento. Nesse sentido, o modelo projetado se fundamenta em três principais etapas conceituais: (i) identificação de novas palavras; (ii) monitoramento temporal do interesse dessas palavras; e (iii) estimativa de potencial de impacto.

Para engendrar essa sequência conceitual delineou-se três tarefas principais de mineração de texto:

(i) Para identificação de novas palavras, tarefa designada como ‘identificação de sinais novos’, utilizou-se como base de conhecimento publicações científicas e o conceito de *embeddings* contextuais dinâmicos de modelos de linguagem pré-treinados capazes de lidar com semântica do texto e a baixa ocorrência das palavras novas, como o BERT;

(ii) Para o monitoramento temporal do interesse das palavras, tarefa designada como ‘identificação de sinais latentes’, foram implementadas técnicas de análise de redes complexas temporais para o monitoramento dinâmico de subgrafos que contém a palavra nova, como o algoritmo *Top-k DBS* apresentado anteriormente. O intuito é identificar o comportamento de crescimento acelerado no entorno da palavra nova, identificando-a como sinal latente. Ou seja, como um nível intermediário entre palavras novas e os sinais fracos, descrevendo palavras que se destacam por algum grau crescente de interesse, mas seriam sinais ainda mais fracos que os sinais fracos, visando distinguir sinais potencialmente relevantes de ruídos;

(iii) Para estimativa de potencial de impacto, tarefa designada como ‘identificação de sinais fracos’, recorre-se à uma segunda base de conhecimento com o objetivo de contrastar os sinais latentes com uma nova fonte de informação a fim de comprovar o potencial desse fragmento de informação identificado como um fragmento de informação relevante, pois propagou-se para outros âmbitos de conhecimento. No caso, sugere-se o uso da base de patentes, realizando uma consulta nos pedidos de patentes a fim de constatar a primeira aparição da nova palavra numa base de conhecimento tecnológico.

Contudo, ao todo, o modelo elaborado como um modelo de descoberta de conhecimento em texto (KDT), abrange mais processos, totalizando 7 (sete) etapas. No Capítulo 4 ‘Modelo Proposto’ encontram-se as justificativas teóricas para cada decisão do processo e na seção 4.3 ‘Demonstração’, apresenta-se uma exemplificação conceitual do modelo, detalhando as etapas e o desdobramento das etapas em subtarefas quando necessário, sugerindo também ferramentas e técnicas para execução. No Capítulo 5 ‘Avaliação Experimental do Modelo’, encontra-se a operacionalização de um protótipo do modelo documentando sua implementação e averiguando quão bem o modelo se sustenta como solução para o problema identificado. Todavia, de modo sucinto, a seguir são elencadas as etapas que estruturam o modelo como solução proposta:

Etapa 1 - Elaborar o Conjunto de Dados;
Etapa 2 - Extrair e Representar Palavras do Conjunto de Dados;
Etapa 3 - Identificar Sinais Novos;
Etapa 4 - Elaborar Rede de Palavras;
Etapa 5 - Identificar Sinais Latentes;
Etapa 6 - Identificar Sinais Fracos;
Etapa 7 - Apresentar Recomendações de Sinais Fracos para Tecnologias Emergentes.

3.3.4 Demonstrar

A etapa de demonstração implica na implementação do artefato em um ambiente controlado, próximo ao real, e tem como foco elucidar o uso do artefato proposto. No Capítulo 4, seção 4.3, consta uma instanciação do modelo, na qual etapas abstratas ganham um exemplo particular (uma instância) representando o passo-a-passo do modelo.

O cenário foi estabelecido a partir do levantamento histórico dos fatos a respeito do grafeno como um caso de tecnologia emergente. Dessa forma, foi possível mapear documentos e datas relevantes para estabelecer a temporalidade para a exemplificação do modelo. A demonstração adota uma sugestão de técnicas a serem implementadas em cada etapa, seguindo o estado da arte para cada tarefa proposta.

3.3.5 Avaliar

A etapa de avaliação tem como finalidade averiguar quão bem o artefato se sustenta como uma solução para o problema. Sob esse viés, essa etapa complementa a etapa demonstração e, diferentemente da anterior que tem como finalidade exibir o uso do artefato, a

etapa de avaliação objetiva a apreciação tanto do rigor da concepção, condução da pesquisa e elaboração da solução, garantindo a validade científica do artefato proposto, quanto da validade funcional do artefato em promover suporte como solução ao problema identificado.

Nesta pespectiva, tendo como guia o Quadro 10, da Seção 3.2, contendo uma coleção de sugestões como processos para avaliação de artefatos concebidos segundo a DSR, adota-se, a avaliação "Descritiva" e "Experimental", as quais recorrem ao 'argumento informado' e a 'simulação'.

A avaliação Descritiva é apresentada em paralelo ao modelo proposto como solução (Capítulo 4, seção 4.2.2 – 'Detalhamento do Modelo'). Informações da base de conhecimento (publicações científicas) são utilizadas para fundamentar e justificar a concepção e utilidade do artefato, argumentando sobre a capacidade conceitual e encadeamento lógico-teórico das etapas que compõem o modelo.

A avaliação Experimental, por sua vez, demonstra a viabilidade e analisa a utilidade da solução a partir da execução do artefato com uma base de dados artificial, construída para esta finalidade. No Capítulo 5, apresenta-se um recorte de um caso real, permitindo comparar, em retrospectiva, se o artefato tem capacidade de antecipar informações sobre tecnologias emergentes. Para a simulação, a tecnologia VoIP (*Voice Over Internet Protocol*), como uma nova técnica de comunicação, foi eleita como cenário para estudo.

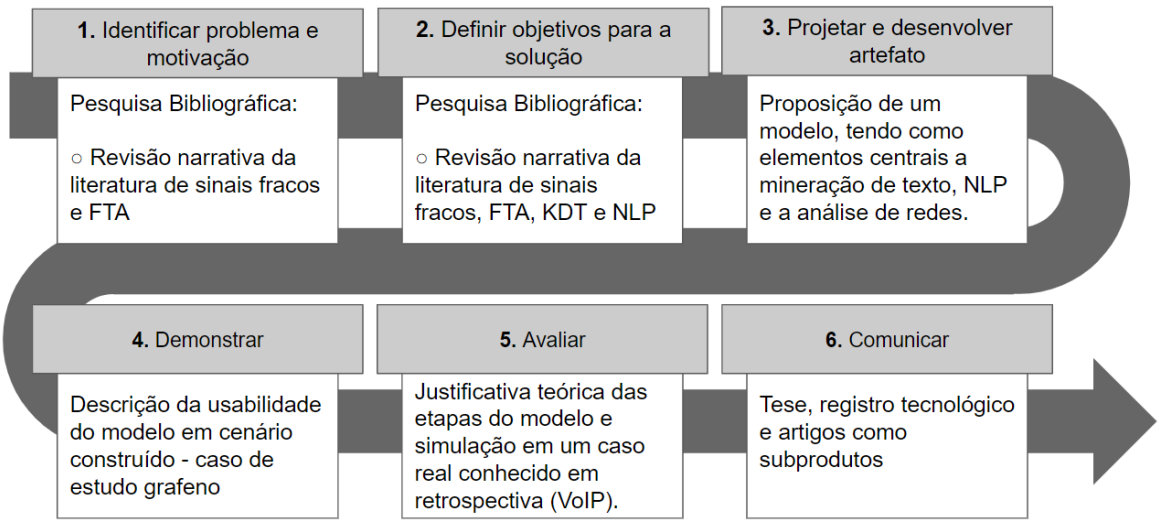
3.3.6 Comunicar

Essa etapa final diz respeito à formalização do conhecimento desenvolvido na pesquisa e decorrente divulgação de sua contribuição. Deve-se enaltecer o problema e sua importância, o artefato, sua utilidade e novidade, o rigor de sua concepção e sua eficácia para pesquisadores e demais público relevante.

Como primeiro produto tem-se a elaboração desta tese. Como demais subprodutos tem-se o registro tecnológico no Instituto Nacional de Propriedade Intelectual (INPI) e a submissão de publicações científicas em congressos e periódicos da área.

Em conclusão, o esquema da Figura 10 resume as práticas envolvidas em cada etapa da DSRM.

Figura 10 – Síntese dos procedimentos e técnicas utilizadas na realização da pesquisa.



Fonte: Elaborado pela autora

4 MODELO PROPOSTO

Este capítulo se encarrega de apresentar o modelo proposto, como artefato resultante da DSRM, a fim de sugerir uma solução para o problema de pesquisa identificado. A primeira seção 4.1 expõe concepções teóricas visando o alinhamento conceitual sobre conceitos-chave para o modelo. Na sequência, a seção 4.2 apresenta o esquema do modelo, ilustrando-o, e também apresenta as justificativas do encadeamento de etapas proposto, descrevendo *por que* foram idealizadas cada uma delas. Por fim, a seção 4.3 se ocupa do detalhamento das diretrizes conceituais que conduzem as etapas do modelo, exemplificando o *que* deve ser realizado.

4.1 CONCEPÇÕES TEÓRICAS

Com o propósito de esclarecer premissas teóricas importantes para concepção do modelo, essa subseção aborda a discussão da compreensão de sinais fracos como palavras novas e o entendimento sobre sinais fracos para tecnologias emergentes.

4.1.1 Palavras Novas como Sinais Fracos

Buscar por novas palavras é coerente ao considerá-las como o menor nível de abstração da informação. Keyes (2021) demonstra, por meio de exemplos históricos, a capacidade de uma palavra exprimir e apontar para um conceito ou ideia. Curioso notar, ainda, como novas palavras podem surgir antes mesmo de um conceito ser completamente compreendido. Conforme exemplifica Keyes, os antigos gregos criaram a palavra ‘*átomo’* para designar as menores e "indivisíveis" partes da matéria, não obstante ao fato, que nenhum grego possuía conhecimento empírico de como realmente era um átomo (não tinham sequer alguma evidência da existência dos átomos). Ao longo do tempo, todavia, o conceito se revelou consideravelmente correto.

Em outro exemplo, menciona os químicos do século XVIII, que insistiam que o fogo dependia de uma substância chamada ‘*phlogiston’*, em contrapartida, no entanto, essa palavra estava associada a uma ideia que se revelou equivocada. Nesse sentido, observa-se que determinados termos científicos podem atuar como mapas bastante confiáveis da realidade, enquanto outros não se sustentam com o tempo. Seguindo esse raciocínio, pode-se notar a capacidade de uma palavra em apontar para um contexto que exprime uma ideia, e assim, inferir sobre a potencialidade da identificação de novas palavras como antecipação de informação, ou seja, compreender uma palavra nova como um sinal fraco.

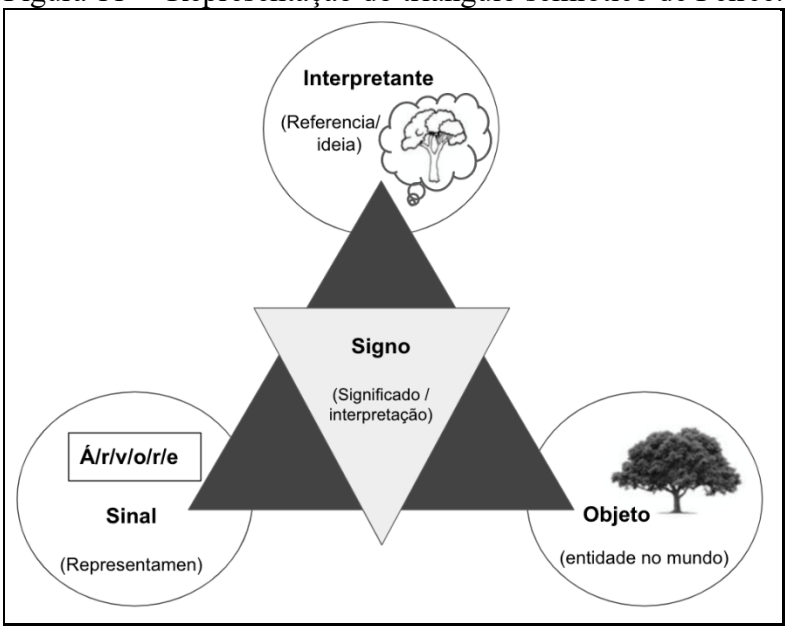
A fim de aprofundar essa discussão, tem-se como embasamento as noções da teoria semiótica de Peirce (SANTAELLA, 2018); o conceito de sinal fraco como fragmento vago de informação de Ansoff (1975); a proposição de signos futuros de Hiltunen (2008) e os estudos de Rossel (2009; 2012); Ahlqvist; Uotila (2020) e Veen e Ortt (2021) pela reflexão entre a existência de sinais fracos independentes (visão ansoffiana) ou dependentes de um processo de interpretação (visão construtivista).

Evidentemente, o conceito de semiótica e a própria teoria semiótica de Peirce (PEIRCE, 1931-1958) são muito mais abrangentes e complexos do que exprime-se nesta tese. Contudo, compreender seu princípio é um importante suporte para a orientação lógica-racional da visão contemplada pelo modelo proposto. Assim, de maneira sucinta, a seguir apresenta-se a ideia central contemplada pelo estudo da semiótica.

A teoria semiótica é o estudo do processo de atribuição e criação de significados num sistema de comunicação por meio de sinais e signos. É um campo de estudo tradicional que apresenta algumas vertentes de pensamento aceitas, sendo uma de suas principais a tríade semiótica proposta por Charles Sanders Peirce[31] (1839 - 1914). A teoria de Peirce enxerga um signo (algo que tem um significado) como uma tríade composta da relação entre: signo, objeto e interpretante (SANTAELLA, 2018). A primeira coisa a se notar é que existem algumas dificuldades terminológicas na compreensão dessa tríade proposta por Peirce, quando ele descreve a tríade que compõe um signo como um dos elementos sendo o próprio signo, isso se torna confuso. O signo ao qual se refere como componente da tríade é o elemento significante (representamen), e não o signo como resultante, como um todo (o signo é maior que o sinal). Isto posto, visando esclarecer o entendimento, assim como Hiltunen (2008), diferencia-se signo como sendo aquilo que surge como resultado da tríade, que "tem significado" e sinal como o componente (representamen), aquele que "aponta para algo". Convém observar que esse intercâmbio de uso das palavras signo e sinal (como representamen) ocorre, pois, num processo de semiose, o signo como resultante, pode ser considerado com um sinal em um ciclo semiótico subsequente. Logo, tem-se o signo explicado pela tríade: sinal, objeto e interpretante (Figura 11).

[31] http://www.peirce.org/

Figura 11 – Representação do triângulo semiótico de Peirce.



Fonte: Elaborado pela autora baseado no triângulo semiótico de Peirce

Diante do exposto, o signo como resultante da tríade depende da relação entre: o sinal (como sendo a forma que o signo assume: imagem, palavra, som, gesto, etc); o objeto (como entidade no mundo: concreta ou abstrata, objeto ou conceito, ao qual o sinal se refere); e o interpretante (como o sentido obtido a partir do sinal, o efeito desse sinal para quem o recebe). Assim, o signo só existe após um processo de significação, surgindo como uma interpretação consensual, uma familiaridade em um determinado contexto. Logo, o signo, como aquilo que existe entre mentes, não é estático. Ao passo que o sinal segue imutável independente de quem o acessa. O sinal é aquilo que promove corpo ao pensamento, é como um estímulo se apresenta.

Com a intenção de tornar os conceitos mais palpáveis, considere como exemplo, um cenário com duas crianças (representando a ausência de um repertório de conhecimento) olhando para o mar. Ambas avistam uma silhueta se aproximando no horizonte, essa imagem desconhecida que surge é um sinal para as crianças. Cada uma tem uma interpretação particular, o sinal tem um efeito diferente na mente de cada uma, a “Criança 1” acha que aquilo se aproximando é um peixe-voador, enquanto a “Criança 2” acha que é um fantasma, uma assombração. Cada uma produz uma ideia diferente a partir de um mesmo estímulo, um mesmo sinal. Enquanto objeto, essa entidade desconhecida que flutua na água retrata o contexto de embarcações. Futuramente, ao compartilharem suas visões poderiam chegar a um consenso, a uma maturidade da ideia originada por aquela silhueta e seu contexto, que tal ‘coisa’ viria a ser definida como uma caravela. Seguindo este raciocínio, este seria o signo, ou seja, quando um

sinal remete a um mesmo conceito para diferentes pessoas. Obviamente este é um exemplo didático bastante simplório para situar o raciocínio entre os conceitos.

Explora-se agora essa explicação para uma palavra como sinal (esquema apresentado na Figura 11). Fazendo uma leitura em paralelo com o exemplo anterior, aqui a grafia da palavra é o sinal, ela remete a uma entidade no mundo, assim como a silhueta era um sinal para o conexto de embarcações. Como interpretante, cada pessoa (cada mente) recorre a uma referência a partir do estímulo do sinal, de modo que quando o sinal é amplamente conhecido, como no caso da palavra árvore, é presumível que a ideia de cada um possua muitas semelhanças entre elas. Mas, à medida que esse sinal decorre de um nicho restrito, específico ou desconhecido, o mesmo sinal, a mesma silhueta, pode gerar ideações tão distantes quanto um fantasma (uma ideia assustadora, negativa) e um peixe-voador (uma ideia entusiasmada, feliz).

O signo, por fim, no exemplo das crianças, seria a caravela, que surge a partir da interação entre mentes, quando existe uma concordância entre os interpretantes e converge em uma noção de compreensão conjunta, um significado. Porque pode-se pensar em várias coisas ao reagir a um sinal, mas só será definido o que ele significa (signo) em conjunto com o contexto de outras mentes. Para a palavra árvore, ela pode ser empregada para além de exprimir a própria entidade árvore, podendo exprimir o conceito de natureza, vida, sustentabilidade, entre outros. Mas, só é possível saber qual dessas associações é factível e definir o signo, ao se saber o que a palavra está significando em acordo com outros interpretantes, em coletivo. O processo de ideação é interno, mas o processo de significação ocorre externamente na interação.

Quando a associação entre algum dos elementos não é evidente para alguém, esta pessoa não consegue resultar no signo. A título de curiosidade, campanhas publicitárias exploram esse recurso de significação, o processo semiótico de estabelecer um signo. Ou seja, para um nicho de pessoas a associação é óbvia, isto é, o significado, um signo, se torna evidente quando se tem clareza sobre os elementos da tríade. Por esse motivo, às vezes, um grupo de pessoas compreende rapidamente a mensagem (o significado) e para outras pessoas é necessário explicar para que compreendam. Isso ocorre porque um grupo conseguiu fazer esse salto de conexão entre sinal, referência e conceito e o outro grupo não. Mas que muitas vezes com o tempo esse signo passa a ser internalizado por mais e mais pessoas tornando-se de fácil entendimento.

Com estes exemplos simples pode parecer difícil de entender esse processo de semiose, pois tudo acontece muito rápido na mente. A todo instante são realizadas conexões e associações que parecem óbvias, entre sinal, objeto e interpretante, mas quando criticamente analisadas existe uma trajetória não-óbvia no processo de significação, criação e transmissão de ideias.

Compreendida a relação entre os elementos que compõem um signo, expande-se a análise adicionando o conceito de sinais fracos na argumentação. Elina Hiltunen (2008), em seu trabalho "*The future sign and its three dimensions*", referência notória entre a bibliografia sobre sinais fracos, exprime o conceito do sinal fraco sob a visão do signo de Peirce. Em sua concepção, o sinal fraco surge como resultado da tríade, nomeando-o de 'signo futuro'. A autora, segundo a categorização proposta por Rossel (2012), tem uma visão construtivista dos sinais fracos, isso implica que, para Hiltunen, o sinal fraco surge a partir do processo de interpretação.

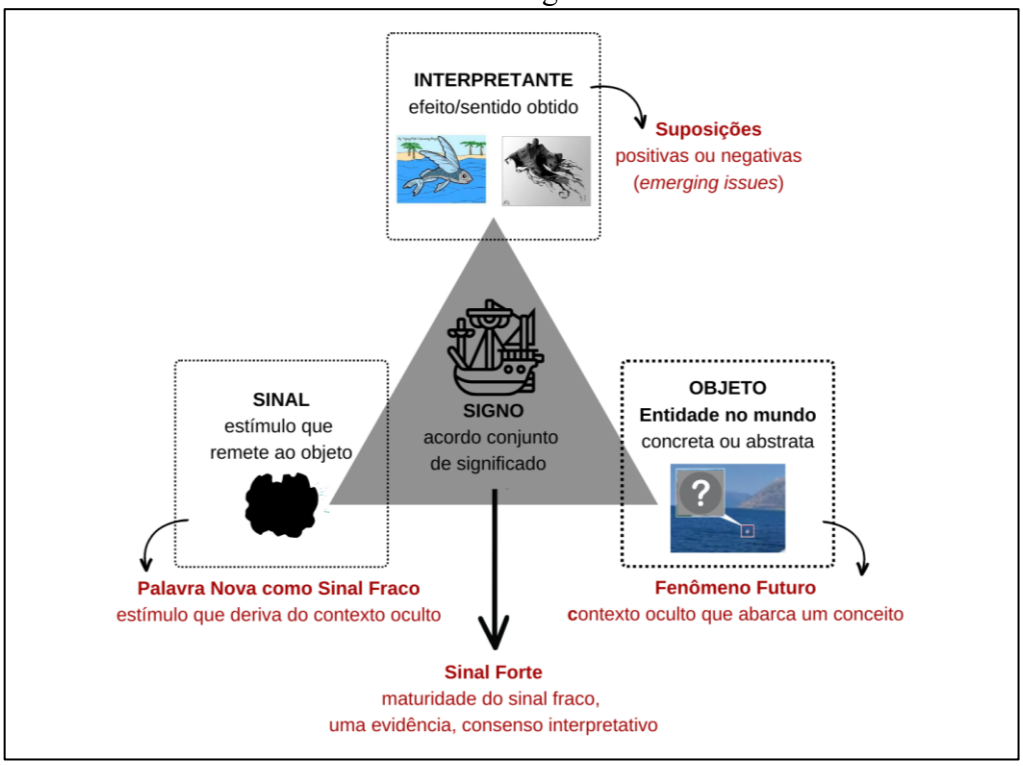
Em contrapartida, esta pesquisa se fundamenta na visão ansoffiana, como colocado por Rossel (2012), na qual compreende-se o sinal fraco como independente de uma interpretação para existir. Tem-se como premissa que, não é porque não se tem a capacidade de prover sentido a um sinal que este não possa existir e ser identificado. O modelo desenvolvido contempla a visão de sinal fraco como sinal (representamen) na lógica de Peirce, em particular, o sinal na forma de palavra.

Seguindo essa linha de raciocínio, pode-se argumentar que a abordagem de Hiltunen cria sinais fracos, buscando observar novas possibilidades de associação entre interpretante e conceito. Enquanto a abordagem proposta tem como objetivo encontrar sinais fracos, buscando identificar palavras novas como o estímulo que deriva de um conceito desconhecido. Tal que a palavra nova encontrada como sinal fraco atua como um mapa para um contexto oculto (objeto) e permite a ideação de suposições positivas e negativas (interpretante).

Nesse sentido, retomando o primeiro exemplo, compreende-se que as silhuetas avistadas seriam como os sinais fracos; as novas palavras descobertas, como indicadores de um evento futuro. O objeto enquanto 'coisa' que flutua é o conceito de embarcação ao qual a silhueta se refere, e seria o contexto oculto ao qual a palavra nova remete (o potencial evento futuro). O efeito dos sinais, o interpretante, peixes-voadores ou assombrações, seriam suposições sobre o possível evento futuro, isso é, potenciais questões emergentes (*emerging issues*). Por fim, a

noção de signo, como consenso são os sinais fortes, que seria a caravela, quando a palavra deixa de ser um indício, um fragmento vago de informação, e se torna um conceito familiar/ conhecido. A Figura 12 e o Quadro 13 trazem uma síntese dessa reflexão.

Figura 12 – Representação da proposição da palavra nova como sinal fraco, unificando os conceitos de Ansoff e do triângulo semiótico de Peirce.



Fonte: Elaborado pela autora (imagens obtidas no Google imagens)

Quadro 13 – A palavra como Sinal de Peirce e Sinais Fracos de Ansoff.

Conceito Peirce		Exemplo	Conceito Ansoff	Exemplo
Sinal	Forma que o signo assume	Silhueta no horizonte	Sinal Fraco (Palavra nova Identificada)	Palavra 'átomo'
Objeto	Entidade no mundo ao qual o sinal se refere	Embarcações ('coisa' que flutua)	Fenômeno Futuro (Contexto Oculto)	Conceito de uma unidade fundamental indivisível da matéria
Interpretante	Sentido obtido a partir do sinal (depende de quem recebe)	Peixe-voador; assombração;	Suposições sobre evento futuro (*emerging issues*)	Potenciais aplicabilidades, efeitos positivos e/ou negativos
Signo	Significado consensual resultante da relação triádica	Caravelas	Sinais Fortes (consenso interpretativo)	Partícula que consiste em um núcleo de prótons e nêutrons, cercado por elétrons

Fonte: Elaborado pela autora

Cabe destacar que a visão de sinal fraco como sinal e não como signo não é excludente. Compreende-se que o conceito de sinal fraco proposto por Ansoff abarca tanto a noção Peirciana de sinal fraco como um sinal (um representamen), onde a existência do sinal fraco é independente da capacidade de interpretá-lo (visão Ansoffiana), quanto como um signo, a

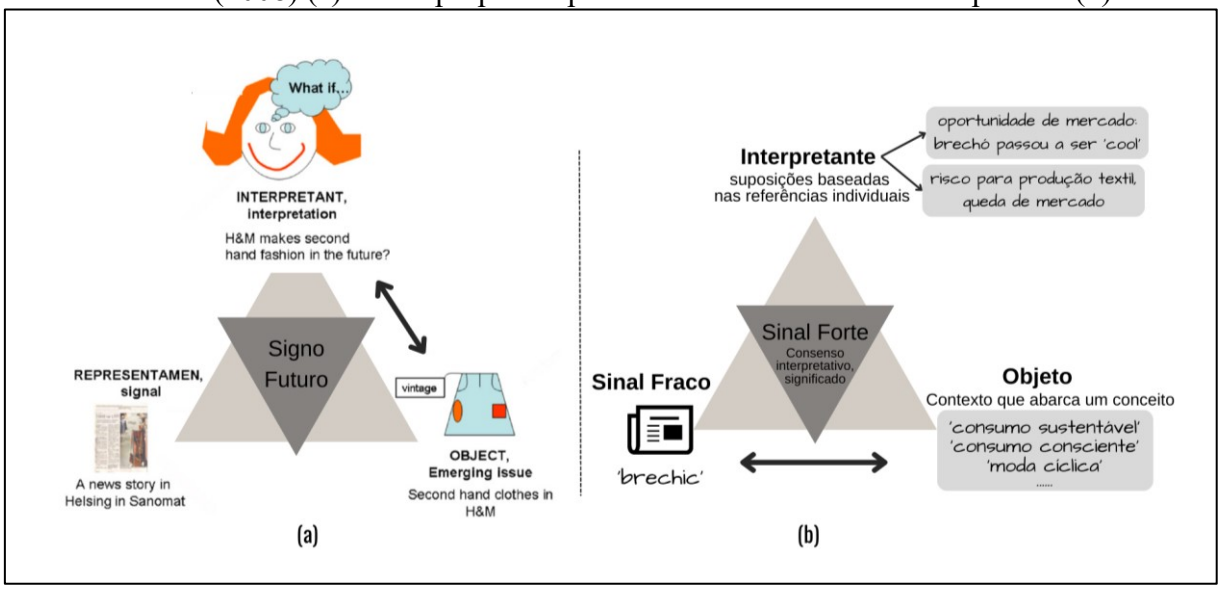
tríade, sendo o sinal fraco como resultado de um processo de interpretação e significação (visão construtivista).

Para ilustrar a diferença, poderia-se dizer que a visão de sinal fraco de Hiltunen está mais voltada para a faceta interpretante da tríade, enquanto a proposta neste trabalho se volta mais para a faceta do objeto. Utilizando o mesmo enredo do exemplo apresentado em seu trabalho, a autora enxerga o sinal fraco como uma nova relação entre 'vintage' e '*fast-fashion'*, suscitando suposições dessa implicação. Como ela mesma pontua, não é o sinal como fragmento de notícia que a fez pensar que é o sinal fraco, pois muitas pessoas podem acessar a mesma notícia e não enxergarem essa nova possibilidade de conexão entre conceitos. Sendo essa percepção, o signo como capacidade interpretativa de enxergar preliminarmente uma possibilidade de significação que é o sinal fraco, o signo futuro, na expectativa que futuramente essa conexão entre conceitos se torne um acordo geral.

Estabelecendo um paralelo, hipoteticamente, a perspectiva aqui apresentada de sinal fraco identificaria a palavra nova '*brechic*' como sinal decorrente de um contexto oculto que enxerga a comercialização de roupas de segunda mão com um novo valor de mercado. Percebe-se que são perspectivas distintas, mas ambas atendem o critério mais importante para definir um sinal fraco: em ser o primeiro sinal de uma possível mudança futura. Como efeito, o primeiro sendo um processo qualitativo para proposição de sinais fracos, dependente da capacidade de interpretação e a segunda, um processo quantitativo baseado na identificação.

A Figura 13 traz uma ilustração das duas percepções, evidenciando que a perspectiva de Hiltunen para um sinal fraco diz mais a respeito sobre a conexão realizada entre a entidade no mundo (objeto) e o referencial de uma pessoa (habilidade de visão), sendo a notícia apenas um aparato de suporte sobre como se entrou em contato com o assunto. Não muito longe, a proposta de sinal fraco como palavra nova refere-se mais sobre a conexão entre a palavra como sinal percebido e o contexto ao qual remete (como entidade no mundo), sendo as questões emergentes consequência da relação entre o sinal fraco e a interpretação sob o referencial individual do especialista que acessar essa informação (interpretante). Nessa proposição, à medida que o objeto, contexto oculto, com o passar do tempo vai se desvendando, o sinal fraco vai transicionando e assumindo características de sinal forte como resultado de consenso dessa tríade. Ou seja, quando uma palavra nova ganha significação coletiva de seu contexto associado e ela deixa de ser uma novidade, um sinal fraco, e tem parte de seu potencial de progresso evolutivo e impacto já revelado.

Figura 13 – Esquema ilustrativo comparando o exemplo de signo futuro apresentado por Hiltunen (2008) (a) com a proposta apresentada de sinal fraco como palavra (b).



Fonte: Elaborado pela autora e adaptado de Hiltunen (2008)

Conclui-se a seção recapitulando que, segundo o conceito semiótico de Peirce, toda palavra é um signo. Isto implica que toda palavra tem um conceito abstrato ao qual um sinal remete. Neste sentido, identificar novas palavras tem potencial para ser considerado um sinal fraco desde que o processo de detecção dessas palavras se encarregue de que a palavra sugerida como sinal fraco tenha seu contexto associado e apresente potencial de impacto, isto é, relevância estratégica. Pois, sob os critérios da teoria de Ansoff, de imediato uma palavra nova atende os critérios de sinal fraco de fragmento vago e ambíguo de informação e também de novidade, bastando assegurar-se da estimativa de relevância estratégica para a palavra nova descoberta poder ser vista como um sinal fraco. Sob essa condição, então, pode-se assumir a descoberta de uma palavra nova, que aponta para um contexto oculto, como primeiro indício de um fenômeno futuro, um sinal fraco. Por fim, destaca-se a relevância da discussão em comparativo com a visão de sinal fraco como signo trazida por Hiltunen (2008), como referência notável para a área, fortalecendo o posicionamento desta tese.

4.1.2 Sinais Fracos para Tecnologias Emergentes

Na investigação da literatura verificou-se a necessidade da identificação de sinais fracos, particularmente, a relação desse tipo de estudo de antecipação de informação para o domínio tecnológico. Mais ainda, pontualmente, o interesse de estudos sobre tecnologias emergentes. Em decorrência disso, esta subseção tem o intuito de apresentar como se compreende termos

importantes para o contexto de desenvolvimento tecnológico como, tecnologia disruptiva, tecnologia emergente, emergência tecnológica e inovação.

De início é preciso salientar, como colocam Warnke e Heimeriks (2008), que o estudo dos desenvolvimentos tecnológicos é uma questão complexa. Tecnologias não advêm da natureza, mas sim de construções realizadas pelo homem; são produtos da evolução cultural. Diante desse cenário, os vários atores envolvidos nesse processo podem usar diferentes definições de tecnologia. Mais que isso, tecnologias estão em constante evolução em um contexto social e suas definições também podem mudar.

Nessa pesquisa entende-se por tecnologia qualquer meio que sirva a um propósito humano resultante da aplicação do saber teórico (método, processo, produto ou material) (CUPANI, 2016). Dito isso, embora o conceito de tecnologia disruptiva seja de extremo interesse para o planejamento estratégico, ela só pode ser reconhecida como disruptiva em retrospectiva (LI; PORTER; SUOMINEN, 2018). A tecnologia emergente, por outro lado, é frequentemente utilizada para descrever a possibilidade de uma mudança dramática e impacto nos sistemas socioeconômicos (COZZENS *et al.*, 2010; ZHU; PORTER, 2002; ROTOLO; HICKS; MARTIN, 2015). Tecnologias emergentes podem representar mudanças incrementais ou radicais (LI; PORTER; SUOMINEN, 2018) e são caracterizadas pela: novidade, persistência, crescimento e formação de comunidade (SUOMINEN; NEWMAN, 2017).

Contudo, nesse momento, julga-se pertinente reforçar como coloca Ansoff, a separação entre um sinal fraco e o fenômeno ao qual se refere. A principal distinção entre uma pesquisa com objetivo em detectar sinais fracos para tecnologias emergentes e um estudo que visa identificar tecnologias emergentes, subsiste no fato de que embora os sinais fracos possam se tornar fortes, nem todos os sinais fracos passam por essa transição. Ao passo que um estudo visando identificar tecnologias emergentes, resultaria em um estudo focado em sinais fortes, pois seria uma tecnologia que de alguma forma já atingiu alguns estágios de maturidade. Sendo assim, uma pesquisa de sinal fraco busca por informações com características de novidade, incerteza e potencial de impacto. De modo que, um sinal fraco para tecnologia emergente antecede a identificação de uma tecnologia emergente. Cabe destacar que ambos processos são válidos e importantes para o desenvolvimento tecnológico, mas com o objetivo de dedicar-se a cada um deles com precisão é necessário defini-los e diferenciá-los.

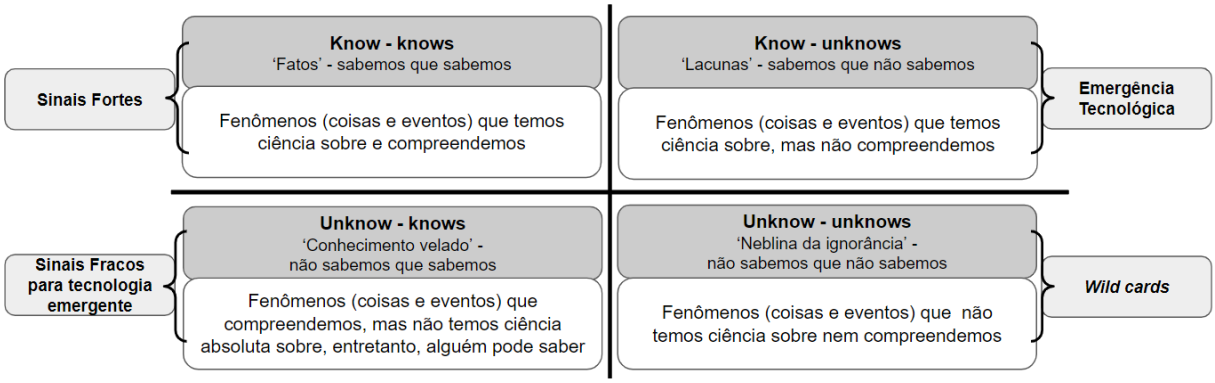
Isso posto, outro ponto a ser destacado, diz respeito aos estudos sobre tecnologia emergente e emergência tecnológica. Muitos estudos empregam as expressões "emergência tecnológica" e "tecnologias emergentes" de maneira intercambiável, apesar de algumas vezes o emprego do termo "emergência tecnológica" no estudo tratar-se de "tecnologias emergentes" e vice-versa (BURMAOGLU *et al.*, 2019; ROTOLO; HICKS; MARTIN, 2015). Visando esclarecer esses rótulos, explicita-se o que se compreende ser um problema de detecção de tecnologias emergentes (busca pela "coisa") e um problema de detecção de emergência tecnológica (busca pela "lacuna"). Em outras palavras, enquanto a identificação de tecnologias emergentes visa descobrir e explicitar informações sobre a tecnologia em si, a identificação de emergência tecnológica objetiva descobrir a ausência de informação que existe entre os artefatos desejados e os recursos científicos/tecnológicos subentendidos, explicitando o que ainda precisa ser desenvolvido para torná-lo possível.

Com o propósito de elucidar ainda mais essa questão, baseado no parecer de Cagnin, Havas e Saritas (2013) sobre um dos desafios do FTA ser distinguir entre "*known-unknowns*", "*unknown-knows*" e "*unknown-unknowns*", apresenta-se uma distinção para "emergência tecnológica" e "tecnologias emergentes" fundamentada no conceito teórico da "Janela de Autoconsciência" de Johari (LUFT; INGHAM, 1955), posteriormente, atribuído para análise de riscos e planejamentos, numa adaptação comumente conhecida por matriz de Rumsfeld[32].

Compreende-se, portanto, a detecção de sinais fracos para tecnologias emergentes, pelo quadrante "*unknow-knows*", revelando os sinais fracos dessas tecnologias antes que se tornem "*know-knows*", ou seja, não se sabe que se saiba, são informações que estão veladas. Ao passo que se enxerga o processo de emergência tecnológica como quadrante de "*know-unknows*" com a descoberta de lacunas na literatura que podem ser preenchidas, mas não se sabe como, ou seja, se sabe o que não se sabe (Figura 14). A fim de complementar a análise, o quadrante '*unknow-unkows*' corresponderia aos *wild cards*, eventos desconhecidos que não se sabe quando irão ocorrer, ou seja, não se sabe que não se sabe, e o quadrante "*kn*ow-*know*", correspondendo aos sinais fortes, informações sobre as quais se tem convicção, sabe-se que se sabe.

[32] Donald Rumsfeld foi Secretário da Defesa dos EUA em 2002.

Figura 14 – Conceitualização dos termos “tecnologia emergente” e “emergência tecnológica”.



Fonte: Elaborado pela autora baseado na matriz de *Johari Window* (1955)

Por fim, diferencia-se o interesse em sinais fracos para tecnologia emergente do conceito de inovação. Entende-se que a inovação é um fenômeno complexo e não depende apenas da descoberta e do desenvolvimento tecnológico, mas também depende de características do mercado e de aceitação, associados à noção de geração de valor. Por essa razão, refere-se a detecção de sinais fracos para tecnologias emergentes e não para inovações (ERZURUMLU, 2017; SHIBATA; KAJIKAWA; SAKATA, 2010).

4.2 APRESENTAÇÃO DO MODELO

4.2.1 Modelo de Detecção de palavras novas como sinais fracos para tecnologias emergentes

Como solução projetada para atender o objetivo da pesquisa “desenvolver um modelo para detectar sinais fracos para tecnologias emergentes considerando métodos e técnicas de descoberta de conhecimento em texto”, tem-se como premissa as concepções teóricas discutidas na seção anterior (4.1) e os objetivos para solução identificados conforme apresentado na seção 3.3.2 (Quadro 12). Na sequência esses achados teóricos são combinados para a proposição de um modelo como solução para o problema de pesquisa.

Face ao contexto de tecnologias emergentes, é pertinente a ideia de identificar novas palavras considerando a relação entre o crescimento de termos técnico-científicos com o desenvolvimento acelerado do conhecimento científico e da produção tecnológica, característica marcante do século XXI (KRIEGER, 2006). O processo de criar novos termos sempre existirá na ciência e tecnologia, como destaca Siegfried (2020), sempre há uma lacuna entre uma palavra e a realidade que ela representa, tal que, uma parte considerável do progresso

científico é estreitar essas distâncias, transformando rótulos vagos em símbolos mais específicos, ou seja, apresentando novas palavras, novos termos.

Logo, o mundo progride e assim também a língua, novas demandas e novas palavras surgem rapidamente. Diante desse cenário, à medida que a tecnologia avança, são necessários termos para se referir à elas. A complementar, como aponta Keyes (2021), o surgimento de novas palavras é imprevisível. Mas, uma vez que as palavras são cunhadas, é possível identificar suas primeiras aparições. Destarte, nesta conjuntura de aceleração tecnológica, buscar por novas palavras condiz com a necessidade de antecipar informações.

Pontua-se ainda, a adequação da proposta de buscar palavras novas como sinais fracos para tecnologias emergentes, em razão de atender a demanda de antecipação de informação, especialmente, suprindo a necessidade da informação "o quanto antes", expressada na literatura como desejo da área de planejamento estratégico. Desta forma, considerando que grande parte dos estudos de planejamento estratégico são centrados em informações sobre tendência, afirma-se a contribuição para a antecipação, para o saber "o quanto antes", a partir da identificação de sinais fracos sobre algo novo (sem precedente), ao descobrir palavras novas, como a menor unidade de informação capaz de apontar para algo, um contexto.

Como meio para operacionalizar essa busca da palavra nova, recorre-se à técnicas de mineração de texto. Isso se sustenta, pois, pela visão ansoffiana, na qual sinais fracos têm propriedades quantificáveis, ou seja, são passíveis de serem extraídos por métodos e técnicas de mineração de texto. Assim, pode-se identificar esses sinais fracos a partir da descoberta de palavras novas como fragmentos de informação textuais presentes em bases de dados textuais, como publicações científicas. Mais que isso, o emprego de técnicas computacionais se justifica por viabilizar a tarefa, considerando o volume de dados disponíveis que impedem a análise eficiente por especialistas, além de minimizar a interferência de viés cognitivo na identificação desses sinais.

A complementar, compreende-se que a palavra completamente isolada não é capaz de fornecer informação suficiente para um processo de planejamento estratégico. Isto é, reconhece-se que o entendimento da palavra não é inerente à própria palavra, mas é criado por sua relação com outras palavras e com o mundo. Nesse sentido, embora uma palavra nova isoladamente não seja capaz de conceder clareza sobre o conceito ao qual remete, por outro lado, a palavra identificada como resultado de um processo de Descoberta de Conhecimento

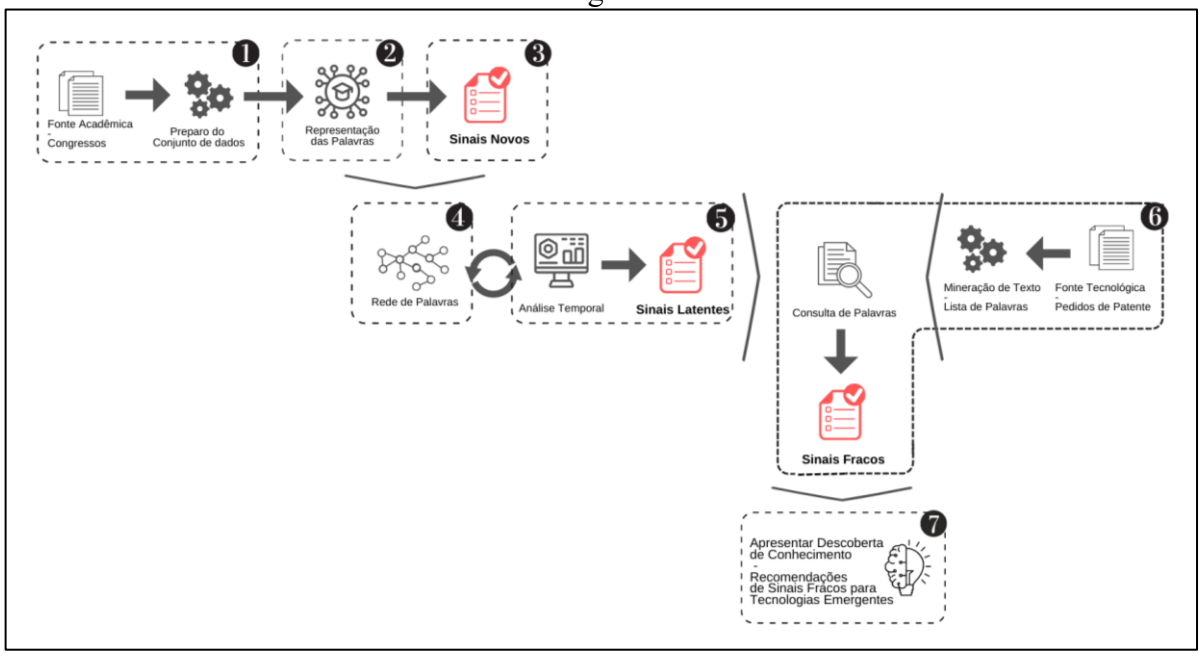
em Texto (KDT) pode ser associada ao seu contexto correspondente, permitindo especular sobre seu entendimento.

No uso de um processo de *embedding,* procedimento intrínseco ao KDT, o contexto da palavra é considerado para sua representação e análise. Logo, ao se identificar a palavra nova, é possível retratar o contexto do qual ela deriva, possibilitando a identificação do conceito ao qual a mesma remete. Justificando assim a detecção de sinais fracos para tecnologias emergentes pela identificação de novas palavras, por meio de métodos e técnicas de KDT, como maneira de antecipar "o quanto antes" informações tecnológicas com potencial relevância para planejamento estratégico.

Finalizando, a compreensão de sinal fraco concebida por Ansoff (1975) tem como pressuposto que, toda ocorrência/fenômeno passa por uma sequência de níveis de conhecimento (do sinal fraco ao sinal forte). O que sustenta a existência de um movimento, uma sequência de eventos que precisa ser capturada com antecedência, à medida que se desenrola (ROSSEL, 2011). Em decorrência disso, o modelo proposto nesta tese tem como base estrutural a lógica da evolução do sinal para sua detecção, assim como também apresentam Griol-Barres *et al*. (2020) em seu modelo.

Considerando todos esses aspectos, sugere-se como solução a identificação do sinal fraco por um encadeamento de etapas, monitorando desde a descoberta da palavra nova até esta ser compreendida como um sinal fraco. A proposta do modelo tem como pilares três principais etapas conceituais. Primeiro, traduz-se a necessidade de antecipação "o quanto antes" dos estudos antecipativos para tecnologia e do planejamento estratégico pela descoberta de novas palavras em publicações científicas. Esta é a etapa voltada à (i) identificação de sinais novos, referindo-se a descoberta de novas palavras. Em seguida, se faz necessário o monitoramento dessas palavras para a distinção entre sinais relevantes e ruídos. Assim realiza-se a etapa (ii), identificação de sinais latentes, como um nível intermediário entre palavras novas e palavras que se destacam por algum grau crescente de interesse, mas seriam sinais ainda mais fracos que os sinais fracos. Por fim, a etapa (iii), identificação dos sinais fracos, é realizada pela estimativa de seu potencial de impacto tecnológico com o uso do conhecimento expresso em uma segunda base de conhecimento, explora-se aqui os pedidos de patentes. No entanto, ao todo, a proposta desta tese elaborada como um modelo de descoberta de conhecimento em texto, apresentado na Figura 15, totaliza 7 etapas que, internamente podem se desdobrar em subtarefas para que sejam realizadas.

Figura 15 – Esquema dos processos do modelo de detecção de sinais fracos para tecnologias emergentes.



Fonte: Elaborado pela autora

A primeira e a segunda etapa se encarregam do preparo da base de dados para tornar a mineração de texto exequível. A terceira etapa é a primeira tarefa de mineração de texto e consiste na descoberta dos sinais novos, as novas palavras. A quarta etapa encarrega-se da elaboração da rede de palavras necessárias para a execução da etapa 5: identificação dos sinais latentes, por meio do monitoramento temporal do comportamento de crescimento acelerado dos sinais novos. A sexta etapa determina quais sinais latentes são considerados sinais fracos para tecnologias emergentes estimando sua relevância estratégica recorrendo à segunda fonte de conhecimento, no caso informações extraídas dos pedidos de patentes. E a última etapa se encarrega do pós-processamento, da recomendação dos sinais fracos, explicitando a descoberta de conhecimento do modelo.

De modo conciso, o modelo é dinâmico e se baseia no monitoramento de três listas de fragmentos de informação textuais selecionados em três estágios diferentes no tempo. Dedicados à busca pelo novo (sinais novos), pelo crescimento (sinais latentes) e pelo potencial de impacto (sinais fracos), temporalmente mantendo e alimentando as listas com as informações que descobre. A seguir é detalhado o encadeamento lógico-teórico das etapas que compõem o modelo.

4.2.2 Detalhamento do Modelo

A concepção do modelo se concentra na busca por novas palavras como sinais fracos para tecnologias emergentes com base em três premissas que se sustentam na literatura: (1) novos termos estão sempre surgindo pela necessidade de explicar novas tecnologias; (2) uma palavra é a menor unidade de informação capaz de exprimir, apontar, para um conceito, uma ideia; e (3) uma palavra nova, identificada por um processo de descoberta de conhecimento, pode ser compreendida como um sinal fraco para o processo de antecipação de informação de planejamento estratégico. Ademais, a ideação da estrutura do modelo como sequência, deriva da própria definição de sinais fracos. A seguir, são detalhadas as justificativas teóricas que sustentam as etapas do modelo.

4.2.2.1 Etapa 1 - Elaborar o Conjunto de Dados

A primeira ação necessária para o modelo diz respeito à elaboração do conjunto de dados sob a qual o modelo raciocina. São necessárias decisões acerca do tipo de informação, onde e como obtê-las, assim como prepará-las e dispô-las.

Para o problema de sinais fracos para tecnologias emergentes, em virtude da relação entre ciência e tecnologia, documentos de publicação científica e de pedidos de patentes foram eleitos como fontes de informação. Decisão corroborada pela revisão sistemática de Mühlroth e Grottke (2018) na qual afirmam que as bases de publicação científica e patentes são as principais fontes de dados na identificação de sinais fracos. O modelo trabalha, então, com dados não estruturados do tipo texto. Esta etapa descreve as decisões acerca do conhecimento científico como fonte para o conjunto de dados, sendo retomada as decisões acerca do conhecimento tecnológico como fonte de dados para o modelo na Etapa 6.

Além de selecionar publicações acadêmicas como fonte de dados, particularmente, apoia-se nas publicações de congressos em ciência e tecnologia (C&T), levando em consideração o comprometimento do modelo com a antecipação e ineditismo da informação. Esse tipo de publicação, embora possa não apresentar o mesmo rigor das publicações de periódicos, representam o que se tem de mais novo na ciência. As conferências costumam ser focadas em novas ideias, de forma que a análise desse conteúdo é uma fonte pertinente de sinais, possibilitando que o "novo conhecimento científico" seja acessado o quanto antes (WATTS; PORTER, 2007; AMANATIDOU *et al.*, 2012; FURUKAWA *et al*., 2015; AMPLAYO; HONG; SONG, 2018; YANG *et al.*, 2022). Em contrapartida, o processo de publicação em

periódicos, que passam pelo processo de "*peer review*", pode levar alguns anos até sua publicação (DERNIS; SQUICCIARINI; PINHO, 2016).

Após selecionar a base de dados sob a qual o modelo irá operar, define-se a granularidade do intervalo de tempo a ser considerado (*timestamp*) (semanal, mensal, anual, semestral, bianual, etc.) para recuperação dos documentos. A granularidade temporal considerada é a anual em razão da frequência com a qual ocorrem as conferências, por conseguinte, a periodicidade com a qual o modelo considera que o conhecimento científico é atualizado. Assim, pode-se obter um conjunto de observações ordenadas no tempo para posterior análise.

Para a extração de informação, foram selecionados como conteúdos relevantes para as tarefas de mineração de texto do modelo, o ano (*year*), o título (*title*) e o resumo (*abstract*). A escolha apenas do resumo como texto principal a ser minerado decorre de algumas razões. Primeiro, o uso do resumo permite abarcar a completude do conhecimento disponibilizado, por ser um trecho de fácil acesso das publicações, já que nem sempre o texto integral está disponível. Além disso, mesmo que a obra esteja em outro idioma, o resumo em inglês é ofertado por ser reconhecida como a língua universal da ciência e da tecnologia (BIDERMAN, 2001). Como segundo ponto, como demostram Tshitoyan *et al*. (2019), o uso do resumo para obter *embeddings* é positivo, pois se tratam de textos concisos com alto valor informacional. Isto evita o emprego de palavras desnecessárias que podem aumentar o ruído na obtenção dos *embeddings,* uma vez que a qualidade e a especificidade do domínio do *corpus* determinam a representatividade dos *embeddings,* acarretando na efetividade das tarefas pretendidas. Por fim, considerando o volume de dados a serem manuseados, utilizar-se dos resumos, e não dos textos em inteiro teor, favorecem a viabilidade do processamento dos dados.

O preparo dos dados, diz respeito ao pré-processamento dos textos. Uma etapa de pré-processamento adequada ajuda a aumentar a precisão das outras tarefas de NLP, por isso a importância desse procedimento. No entanto, a escolha das tarefas a serem executadas depende do objetivo do modelo e da escolha de técnicas de representação vetorial do texto a serem implementadas. Usualmente, pode-se encontrar como preparo dos dados tarefas de análise léxica, convertendo o texto original em uma lista de palavras (*token*ização), bem como tratar pontuações, hífens, dígitos e também uniformizar letras maiúsculas e minúsculas. Outras tarefas incluem, ainda, a radicalização (*stemming*) ou lematização (*lemmatization*) das palavras, implementação de *stop-words* ou *stop-lists* e tarefas de padronização, como representação de

erros de digitação e expansão de contrações, visando abranger variações de escrita (ALLAHYARI *et al*., 2017).

Como resultado desta etapa tem-se disponível um conjunto de dados de textos semanticamente densos (resumos) com informações atuais sobre ciência e tecnologia de um determinado ano (conferências). Por fim, para o armazenamento e a disponibilização dos dados, bancos de dados orientados a documentos que permitam consultas e atualização dos documentos, se mostram adequados.

4.2.2.2 Etapa 2 - Extrair e Representar Palavras do Conjunto de Dados

Essa etapa concerne a escolha adequada da técnica de representação vetorial do texto para realização das etapas subsequentes de mineração. Para processamento de dados não estruturados, como texto, é preciso abordagens de NLP para representação das palavras. Como visto na seção 2.4.1, técnicas de *word embedding* é o nome atribuído ao processo de codificação das relações entre palavras por meio de representações vetoriais. Especificamente, considerando o objetivo do modelo, torna-se necessária uma representação vetorial semântica contextual dinâmica das palavras. Isto é, técnicas capazes de gerar *embeddings* diferentes para uma mesma palavra que apresente contextos distintos, ou seja, que a representação vetorial seja análoga ao significado e sentido da palavra.

Atualmente, tal representação pode ser obtida com modelos de linguagem pré-treinados, como ELMo® (PETERS *et al*., 2018), UMLFiT® (HOWARD; RUDER, 2018), BERT (DEVLIN *et al.*, 2019), GPT-2® (RADFORD *et al.*, 2019), GPT-3® (BROWN *et al.*,2020) entre outros. Todavia, o modelo BERT se destaca na capacidade em obter *embeddings* ao nível de palavras (*language understanding task - bidirectional encoder-only transformer architecture*) se comparado com o GPT, por exemplo, melhor na geração de textos (*text generation task - unidirectional decoder-only transformer architecture*). Além de ser capaz de lidar com questões de palavras OOV (*out-of-vocabulary*) (KHODAK *et al.*, 2018; HU *et al.*, 2019) e palavras raras (SCHICK; SCHÜTZE, 2020; LI, X. *et al.*, 2021), essencial para a etapa seguinte de descoberta de novas palavras. Por essas razões o modelo proposto toma como base o uso do modelo BERT.

A entrada desta etapa é constituída pelo conjunto de dados e como saída tem-se os *embeddings* das palavras considerando sua semântica contextual dinâmica.

4.2.2.3 Etapa 3 - Identificar os Sinais Novos

Identificar sinais novos no modelo é sinônimo de descobrir palavras novas. Podem ser consideradas como palavras novas tanto palavras ou expressões novas no sentido de neologismo lexical ou do neologismo conceitual, quando uma palavra já existente adquire novo sentido, como visto na seção 2.4. De modo que podem ser compreendidos como dois subproblemas paralelos, um que compreende a identificação de novas palavras inéditas e o problema de identificação de mudança de contexto de palavras existentes, problema conhecido também na literatura de NLP por *semantic shift* (KUTUZOV *et al.*, 2018).

No entanto, independentemente do caso, detectar palavras novas e obter suas representações vetoriais por meio de técnicas de NLP apresentam-se como um desafio, tendo em vista que palavras novas estão fora do vocabulário dos modelos pré-treinados e/ou apresentam baixa frequência de ocorrência nos conjuntos de dados. Assim sendo, esta é uma discussão bastante recente sobre como inserir/ensinar novas palavras para um modelo de *embeddings* de aprendizado de máquina (HOMBAIAH *et al.*, 2021).

No porvir, com a evolução dos modelos LLMs (*Large Language Models*), seguramente se aperfeiçoará a capacidade das abordagens em identificar palavras novas e, assim, as decisões desta etapa irão se adaptar às técnicas de NLP disponíveis. Não obstante, na seção 4.3.3 investigaram-se abordagens para lidar com palavras OOV e raras e sugere-se uma rotina para identificar palavras novas (inéditas).

Pontua-se ainda que o modelo assume para seu racional apenas palavras simples. Isto é, toma como premissa que palavras compostas são indentificadas pelo modelo, visto que adota-se como parte do processo de identificação a análise das redes de palavras. Assim, palavras compostas são identificadas a partir da detecção do caso de *semantic shift* em conjunto com a representação em rede do contexto da palavra, dado que muitos termos tecnológicos tomam emprestado palavras ou conceitos existentes para compor sua denominação. De modo que termos como "*machine learning*", "*quantum computing*" ou "*artificial intelligence*", por exemplo, seriam capturados pela mudança de contexo de emprego de alguma das palavras que compõem a nomenclatura da nova tecnologia. No caso de duas palavras inéditas compondo o termo, são identificadas pela detecção do ineditismo e análise da representação contextual da rede de palavras.

As novas palavras identificadas recebem a classificação de ‘sinal novo’ e como resultado desta etapa tem-se uma lista de ‘sinais novos’ armazenando as palavras novas respectivas ao seu ano de referência.

4.2.2.4 Etapa 4 - Elaborar a Rede de Palavras

A detecção de sinais fracos é um processo contínuo de varredura em busca de pistas que indiquem se mudanças discretas estão ocorrendo. Logo, a escolha do uso de Ciência de Redes para o modelo se destaca por permitir a inserção contínua de novas palavras na estrutura, conferindo dinamicidade e reaproveitamento do modelo, questão apontada na literatura como uma fraqueza de modelos anteriores (MÜLHROTH; GROTTKE, 2018). Pois, métodos que não permitem atualizações no *corpus* tornam a tarefa de detecção custosa por requerer que o processo de mineração de dados seja iniciado do zero.

O uso de redes é uma abordagem bastante tradicional para mapear a ciência investigando redes de citações entre publicações ou autores. Entretanto, é difícil para os métodos baseados em citações conseguirem detectar aquilo que é mais recente porque os artigos geralmente requerem tempo para serem citados por outros artigos. Em contrapartida, redes de palavras podem ser construídas tão rapidamente quanto novos artigos são publicados (KATSURAI; ONU, 2019). A complementar, análises baseadas em dados bibliométricos (como, autores, citações, palavras-chave), não consideram o texto em si (a parte não estruturada da informação), deixando fora da análise grande parte do conhecimento que pode estar oculto nos dados textuais. Assim, a rede de palavras proposta, baseada na semântica das palavras novas identificadas e seu contexto, apresenta capacidade de descobrir tecnologias emergentes com mais antecedência se comparado com redes baseadas em dados bibliométricos.

Pela hipótese distribucional, palavras têm significados semelhantes quando são usadas em contextos semelhantes. Em razão disso, contextos semelhantes geram vetores semelhantes, ou seja, mais próximos. Portanto, justifica-se construir a rede com os vetores de palavras e visualizá-la em um grafo. O processo de *embeddings* em si posiciona as palavras em um espaço vetorial, de forma que palavras em contextos semelhantes estejam próximas. Todavia, não é possível saber o quão relacionadas estão e, dessa forma, é estabelecido um critério de relação entre as palavras. Para construir a rede, as palavras (seus *embeddings*) foram determinadas como nós e a intensidade de suas relações obtida pelo cálculo de similaridade como arestas.

Além disso, a visualização da rede se mostra adequada, pois como ressalta Moro *et al.* (2018), tecnologias emergentes são intrinsecamente complexas e podem, às vezes, não serem identificadas com uma única palavra. Neste sentido, visualizar a relação da palavra nova com as demais palavras que compõem seu contexto pode facilitar a identificação de um conceito mais complexo e identificar uma nova tecnologia emergente.

Como resultado desta etapa produz-se a rede de palavras nas quais se inserem a palavra nova identificada, representando sua interação com as demais palavras do contexto.

4.2.2.5 Etapa 5 - Identificar os Sinais Latentes

Considerando a característica evolutiva dos sinais fracos, os sinais novos precisam ser monitorados para considerar sua plausibilidade, distinguindo-os dos ruídos, configurando uma classe intermediária de sinais denominada de sinais latentes.

A antecipação de informações, com a detecção de sinais fracos, procura investigar fragmentos de informação que se despontam. Grande parte dos estudos determinam sinais fracos como uma classe de termos com uma taxa média de crescimento anual relativamente alta e uma frequência de ocorrência relativamente baixa. Dessa forma, buscou-se inserir no modelo uma forma dinâmica de monitoramento capaz de realizar atualizações ao longo do tempo. À vista disso, técnicas de Análise de Redes Complexas Temporais se mostram adequadas, possibilitando acompanhar temporalmente os sinais novos identificados, e assim, permitindo distinguir aqueles que são potenciais sinais fracos (sinais latentes) daqueles que são apenas ruídos.

A nomenclatura ‘sinal latente’ foi inspirada no emprego da expressão nos trabalhos de Park e Cho (2017) e Lee e Park (2018). Em seus estudos o termo foi utilizado para se referir ao terceiro quadrante do método proposto por Yoon (2012). Embora, o modelo proposto não derive da mesma abordagem, o termo em seu sentido conceitual é empregado para se referir a um sinal intermediário, que ainda não é significativamente perceptível, isto é, um sinal ainda mais vago e incipiente que precede um sinal fraco, podendo tanto vir a ser um sinal fraco quanto a não apresentar potencial de impacto e cair no esquecimento. Como exemplo, pode-se pensar em neologismos e jargões, que são mencionados apenas uma ou outra vez em publicações científicas (BARAM-TSABARI *et al*., 2020). Ou seja, é importante analisar o crescimento das relações das novas palavras como indicativo de interesse na palavra ao longo do tempo e,

eliminar, dentre aquelas palavras identificadas como novas, o que pode ser considerado apenas ruído.

Para poder classificar uma palavra como ‘sinal latente’, técnicas que possibilitam monitorar a dinamicidade temporal da rede e identificar um comportamento de crescimento utilizam-se de algoritmos de análise de redes temporais para identificação de subgrafos ou comunidades. Particularmente, a análise *burst* é candidata em satisfazer essa necessidade, capaz de identificar uma densidade acumulada rapidamente, um crescimento na intensidade da relação entre um subconjunto de nós da rede, em um determinado intervalo de tempo. Dernis, Squicciarini e Pinho (2016) trazem o conceito da análise *burst* para investigar bases de documentos de publicação científica e patentes a fim de monitorar o desenvolvimento tecnológico. Katsurai e Ono (2019) também fazem uso da métrica *burst* para investigar o comportamento das redes de palavras de publicações científicas. Ressalta-se que modelos que implementam o conceito de *burst* permitem mensurar não apenas o crescimento, mas a velocidade de crescimento das relações na rede. Dessa forma, se mostra uma métrica relevante quando se está interessado em identificar uma alteração “o quanto antes”.

Como resultado desta etapa tem-se uma lista de palavras denominadas sinais latentes. Onde um sinal novo é classificado como um sinal latente quando um subgrafo de saída do algoritmo de análise *burst* inclui um nó referente a palavra nova presente na lista de sinal novo. Então, a saída como lista de sinais latentes corresponde às palavras novas que apresentam rápido crescimento na intensificação de suas relações contextuais, sendo candidatas a sinais fracos.

4.2.2.6 Etapa 6 - Identificar Sinais Fracos

Esta é a etapa principal do modelo, na qual ocorre a convergência das novas palavras identificadas (sinais novos) e classificadas como sinais latentes, em sinais fracos. Pela definição, um sinal fraco se caracteriza por sua novidade, sua incerteza (fragmento vago) e seu potencial de impacto (relevância estratégica). Dessa maneira, objetiva-se associar os sinais latentes aos potenciais impactos tecnológicos para que sejam considerados sinais fracos para tecnologias emergentes.

Nesse momento recorre-se à uma segunda fonte de conhecimento a fim de comprovar a relevância da palavra nova identifica como sinal latente até então. O confrontamento com uma segunda base de informação tem a pretensão de estimar a relevância do sinal averiguando que

a palavra nova ocorre também em outras bases de conhecimento relevantes para o domínio de interesse, agindo como um meio de estimar o potencial de impacto desse sinal identificado.

Para o modelo desta tese, argumenta-se sobre o uso da base de patentes como fonte de conhecimento tecnológico. Precisamente, tem-se interesse no registro de pedido de patente, visto que se busca antecipar informações sobre tecnologias emergentes e visto que pode haver um lapso de tempo entre o pedido e a concessão da patente. Conforme o racional do modelo, não há relevância se a patente foi ou não concedida, o ponto de interesse está na aparição sem precedente da palavra na base de conhecimento tecnológico. Para o modelo, essa aparição da palavra nova nessa segunda base de conhecimento é indício de projeção tecnológica, isto é, uma estimativa de relevância estratégica e potencial impacto.

A estrutura dessa etapa tem como base dois pontos principais: a premissa de que a ciência ocorre antes da tecnologia, como corroboram Wu *et al.* (2010). Sendo assim, os autores postularam, portanto, que os primeiros sinais são encontrados inicialmente em publicações científicas e apenas mais tarde em bancos de dados de patentes. E, segundo, o fato das patentes serem consideradas as melhores fontes de informação tecnológica (ERNST, 2003). O autor menciona que até 80% do conhecimento tecnológico pode ser considerado como implícito em patentes. Sob essa ótica pode ser considerada uma base sólida para o monitoramento da tecnologia (TEICHERT; MITTERMAYER, 2002).

Ademais, embora campos tecnológicos apresentem graus diferentes de dependência da ciência, a maior parte das patentes relaciona-se com a ciência e conta com citações acadêmicas (WANG; LI, 2021). Tal fato, sustenta a proposição de comparar palavras extraídas de publicações científicas com a base de patentes. A complementar, essa abordagem se mostra promissora considerando os resultados de Raan (2017) e Winnik (2018) que investigam as relações de publicações *'sleeping beauty'* e patentes, indicando que publicações com revelações tecnológicas notáveis costumam ter seus conhecimentos primeiro retratados na base de patentes do que disseminados em citações de publicações em periódicos.

Adicionalmente, o modelo considera, como apontado por Mühlroth e Grottke (2018) e Zhao, Tang e He (2023), a necessidade de se inserir mais de um tipo de base de dados para detecção de sinais fracos. Respeitando, como pontuam Rousseu, Camara e Kotzinos (2021), que as bases escolhidas devam ser coerentes entre si, de modo a serem complementares em fornecer informações adicionais para que seja possível destacar informações relevantes. Sendo

assim, a base de patentes pode ser vista como complementar a base de publicações científicas, visto que patentes citam artigos científicos para embasar os registros tecnológicos, seguindo a lógica do modelo de inovação linear tradiconal, onde o conhecimento flui da ciência básica para a aplicação técnica.

No mais, tendo em vista que nem todo desenvolvimento tecnológico é patenteado, pode-se complementar essa base de conhecimento tecnológico agregando documentos de relatórios técnicos ou de consultorias e revistas ou *websites* renomados na área de novidades tecnológicas. Visando ampliar essa estimativa de projeção tecnológica como estimativa de relevância estratégica e potencial de impacto.

Como consideração final, menciona-se que não seria suficiente, como sinal fraco, identificar apenas a primeira aparição de uma palavra nova somente na base de patentes. Primeiro, textos da base de patentes são mais propensos a apresentarem jargões, palavras novas que ocorrem raríssimas vezes, devido ao processo de tentar nomear algo novo. Nesse sentido, o modelo proposto se encarrega de averiguar o desdobramento da influência da palavra nova antes de consultar a base de palavras tecnológicas. Além disso, a proposição do modelo considera sua possível generalização para outros domínios, como mídias sociais, por exemplo, de forma que essa arquitetura do modelo em três etapas centrais, identificando, acompanhando e averiguando se mostra pertinente.

Como resultado desta etapa descobre-se uma palavra nova como recomendação de sinal fraco para tecnologias emergentes.

4.2.2.7 Etapa 7 - Apresentar Recomendações de Sinais Fracos para Tecnologias Emergentes

Compreende-se que muitas vezes a palavra isolada não é suficiente para contribuir com novos *insights* para planejamento estratégico. Sob essa óptica, as informações descobertas pelo modelo, como etapa de identificação, precisam ser divulgadas a fim de fomentar processos de análise e interpretação. Por essa razão, como última etapa do modelo são elencadas, principalmente, duas formas para a apresentação da informação. Embora a visualização da informação não seja foco deste trabalho, em virtude da profundidade a qual pode se estender este campo de estudo, julga-se necessário pontuar a importância da visualização do domínio tecnológico ao qual o sinal fraco detectado corresponde, ou também, a evolução temporal dos sinais encontrados.

A identificação da palavra nova atua como um guia para investigar um evento futuro que o sinal fraco anuncia. O sinal fraco enquanto possíveis tecnologias emergentes, podem ser tanto um conceito que consegue ser expresso pontualmente por uma palavra, ou por um conjunto de palavras. Significa dizer que, uma palavra isoladamente pode não coincidir com o nome de uma tecnologia emergente, mas estar estreitamente relacionada às palavras que a descrevem. Por esse motivo, sugere-se a visualização do domínio de ocorrência do sinal fraco detectado com posterior análise do contexto correspondente.

Considerando a rede de palavras elaborada para o monitoramento temporal, pode-se resgatá-la e realizar um recorte do subgrafo que contém o sinal fraco para sua visualização, possibilitando encontrar significado em um grupo de palavras que compartilham um contexto, em vez de uma palavra individual. Etapa essencial para a sugestão de possíveis tecnologias emergentes que são apontadas por sinais fracos advindos de ocorrências de *semantic shift* ou descritos por palavras compostas como mencionado anteriormente.

Propõem-se também, uma visualização da evolução temporal das palavras a partir das listas de sinais novos, sinais latentes e sinais fracos, permitindo visualizar quais sinais ocorreram concomitantemente em cada ano, bem como a evolução de seus subgrafos ao longo do tempo. Essa informação sobre a trajetória temporal pode ser relevante para gerar *insights*.

Por fim, salienta-se que a detecção de sinais fracos não é um fim, não são previsões explícitas, ela é um início e tem a exata intenção de ser um gatilho para instigar a reflexão e a busca por informações adicionais em um processo subsequente de interpretação “do desconhecido” o qual os sinais fracos apontam.

Para concluir essa seção, o Quadro 14 sintetiza as etapas e as tarefas envolvidas para operacionalizar a demonstração do modelo que será explorada na seção 4.3 seguinte.

Quadro 14 – Síntese das atividades do modelo proposto.

Etapa		Descrição da Tarefa
1	Elaborar o Conjunto de Dados	1.1 Recuperar e extrair informação; 1.2 Pré-processar dados; 1.3 Disponibilizar *corpus*.
2	Extrair e Representar Palavras do Conjunto de Dados	2.1 Representar computacional e semanticamente as palavras (*embedding* contextual dinâmico).
3	Identificar Sinais Novos	3.1 Identificar palavras novas; 3.2 Classificar como Sinais Novos.
4	Elaborar Rede de Palavras	4.1 Calcular de Similaridade; 4.2 Exibir Rede.
5	Identificar Sinais Latentes	5.1 Monitorar temporalmente o crescimento dos sinais novos; 5.2 Classificar como Sinais Latentes.
6	Identificar Sinais Fracos	6.1 Estimar impacto dos Sinais Latentes; 6.2 Classificar como Sinal Fraco.
7	Apresentar Recomendações de Sinais Fracos para Tecnologias Emergentes	7.1 Visualizar Domínio Tecnológico e/ou temporalidade dos Sinais.

Fonte: Elaborado pela autora

4.3 DEMONSTRAÇÃO

Esta seção retrata uma instanciação do modelo demonstrando conceitualmente sua operacionalização por meio de algumas ferramentas e métodos alinhados às necessidades justificadas na seção anterior (4.2). A exemplificação ocorre em um cenário controlado elaborado com a finalidade de ilustrar e elucidar o desenrolar das etapas do modelo. Destaca-se, que levando em conta a complexidade do modelo e o volume de etapas, a fim de apresentar uma descrição clara, o foco reside apenas no caso da identificação de uma palavra nova como inédita. Dito isso, como elemento para estudo, a palavra ‘grafeno’ foi escolhida como referente a uma tecnologia emergente.

O grafeno é um material de grande interesse para a ciência e tecnologia devido às suas propriedades e potenciais aplicações. O grafeno é um material 2D e pode ser descrito como uma folha bidimensional de átomos de carbono, densamente compactados e dispostos em uma única camada, tendo sua estrutura de cristais semelhante a uma colmeia (RANDVIIR; BROWSON; BANKS, 2014). É o material mais leve, mais forte e mais fino conhecido, além de ser o melhor condutor de calor e eletricidade já descoberto (DREYER; RUOFF; BIELAWSKI, 2010; PATENT INFOTMATICS TEAM, 2011). Ante essas informações, torna-

se evidente o potencial tecnológico desse material, que até o momento presente, é considerado ainda em estágio embrionário de exploração para o mercado.

A história do grafeno possui como marco inicial o trabalho de Phillip Wallace de 1947, por tratar da descrição teórica das propriedades elétricas do grafite. Entretanto, ele não emprega o termo "grafeno", mas refere-se a ele pela expressão "camada hexagonal única" (expressão original, *"single hexagonal layer"*). Naquele momento, não se achava possível um material em 2D existir livre ou isolado "*isolated or free-standing graphene*". Mas, em 1962, dois químicos, Ulrich Hofmam e Hanns-Peter Boehm, observaram flocos extremamente finos de grafite e, em 1986, cunharam o termo 'grafeno' para diferenciar o material *single-layer* (2D) do grafite (3D).

Em 1997, a IUPAC[33] formalizou a nomenclatura do grafeno incorporando-a em seu Compêndio de Tecnologia Química. Mas, somente em 2004 (42 anos após ter sido observado pela primeira vez e 57 anos após sua teorização), Andre Geim e Konstantin Novoselov extraíram cristalitos de um átomo de espessura de grafite, isto é, cristais 2D. Ou seja, conseguiram obter grafeno e demonstraram empiricamente suas propriedades elétricas, conferindo a eles o prêmio Nobel de Física de 2010 (RANDVIIR; BROWSON; BANKS, 2014; DREYER; RUOFF; BIELAWSKI, 2010; PATENT INFOTMATICS TEAM, 2011).

Nesse hiato, entre a proposição do termo "grafeno" (1962) e a publicação histórica de Geim e Novoselov de 2004, o termo grafeno podia ser encontrado também em estudos no contexto de nanotubos de carbono, o que, de modo simplório, são "folhas de grafeno enroladas" (RANDVIIR; BROWSON; BANKS, 2014). Mas, até 2004 não haviam conseguido estabilizar uma folha única de grafeno, por isso a importância do estudo de 2004 de Geim e Novoselov.

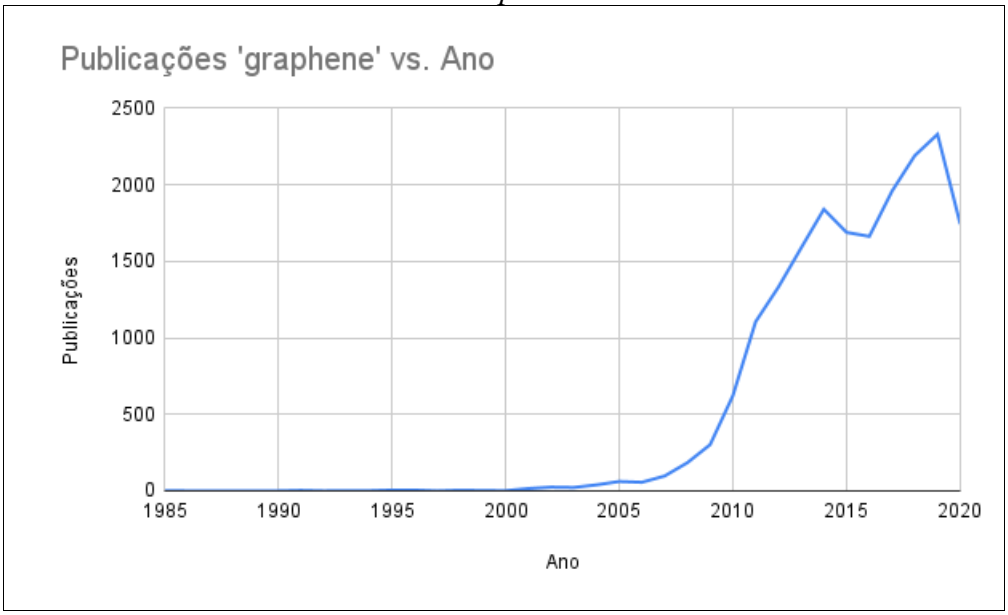
A primeira menção ao grafeno aparece em uma patente (US5376450 A) publicada em 12 de dezembro 1994, tendo prioridade desde 1991, atribuído a UCAR[34] (*Carbon Technology Corporation*) (PATENT INFOTMATICS TEAM, 2011). Convém observar que o documento não menciona a palavra *'graphene*' em seu resumo, menciona apenas uma vez a palavra no corpo do texto na seção de descrição.

A Figura 16 apresenta o gráfico da quantidade de publicações em congressos nas quais a palavra '*graphene*' está presente no título, resumo ou palavras-chave (as informações dizem

[33] *International Union of Pure and Applied Chemistry.*
[34] Em 2002, a empresa mudou seu nome de UCAR para Graftech.

respeito apenas à base *Scopus*). São quase 20 mil publicações até a década atual, contabilizadas a partir da publicação mais antiga, de 1985.

Figura 16 – Gráfico de publicações em conferências a partir da palavra '*graphene*' na base *Scopus*.



Fonte: Elaborado pela autora

Pela tendência apresentada no gráfico é possível traçar um paralelo com as datas importantes na história do grafeno. A primeira publicação é de 1985, consoante ao surgimento do termo grafeno; a partir de 2003 o número de publicações começa a apresentar crescimento significativo, de acordo com o período em que Geim e Novoselov conseguiram extrair os cristais 2D de grafeno. Por fim, a partir de 2010 o número de publicações explode acompanhando o período da publicação do prêmio Nobel, indo de cerca de 500 publicações em 2010, para mais de 2 mil em 2019.

Tendo em vista os marcos da pesquisa do grafeno e o conhecimento do processo do modelo, pode-se apurar se as etapas do modelo proposto estão de acordo. A expectativa não é identificar pontualmente esses trabalhos mencionados na cronologia da história do grafeno, mas usar esses marcos para situar a sequência de etapas do modelo. Estima-se que por volta de 2000 teria sido possível sugerir a palavra grafeno como um sinal fraco, conforme o início de seu crescimento nas publicações. De modo que, uma década antes da palavra se tornar uma tendência (*buzzword*), ela seria recomendada como um sinal fraco para tecnologias emergentes.

Na sequência são pormenorizadas as etapas do modelo apresentado no Capítulo 4. Para exemplificar a demonstração, tem-se como base as publicações em congressos disponibilizadas na base *Scopus* a partir da busca manual da palavra *'graphene'* como descritor.

4.3.1 Etapa 1 - Elaborar o Conjunto de Dados

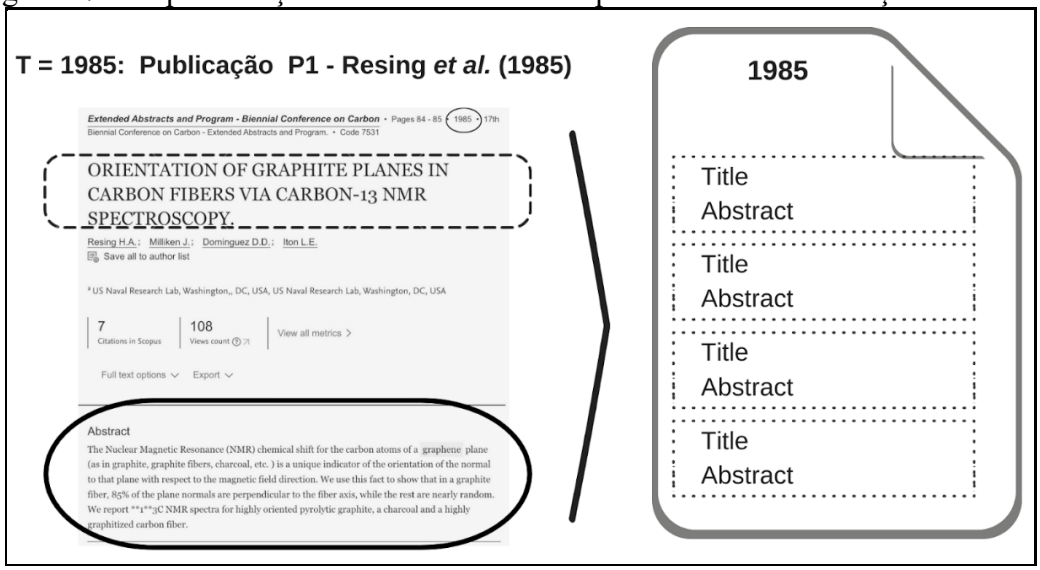
A primeira etapa é responsável pelas subtarefas de (1.1) coleta dos documentos, (1.2) preparo dos dados e (1.3) disponibilização do conjunto de textos utilizado nas etapas subsequentes do modelo.

(1.1) Recuperação e Extração da Informação

Para a demonstração supõe-se que o momento de análise seja janeiro de 1986. Dessa forma, documentos referentes às publicações em congressos sobre ciência e tecnologia no ano de 1985 são recuperados das principais bases de publicação científica como *Scopus*, *Web of Science*, *IEEE*, *ACM*, *Spring Link* e outras. Durante a coleta, para manter organizado os documentos recuperados, são eliminados arquivos duplicados ou inconsistentes, isto é, documentos que não apresentem alguma informação de relevância para a análise, como ausência de data, título ou resumo.

O conjunto de dados é elaborado extraindo os campos, ano (*year*), título (*title*) e resumo (*abstract*) dos documentos recuperados. Dentre eles estaria a publicação de Resing *et al*. (1985), a qual menciona, pela primeira vez, a palavra '*graphene*' em um resumo, como mostra a Figura 17.

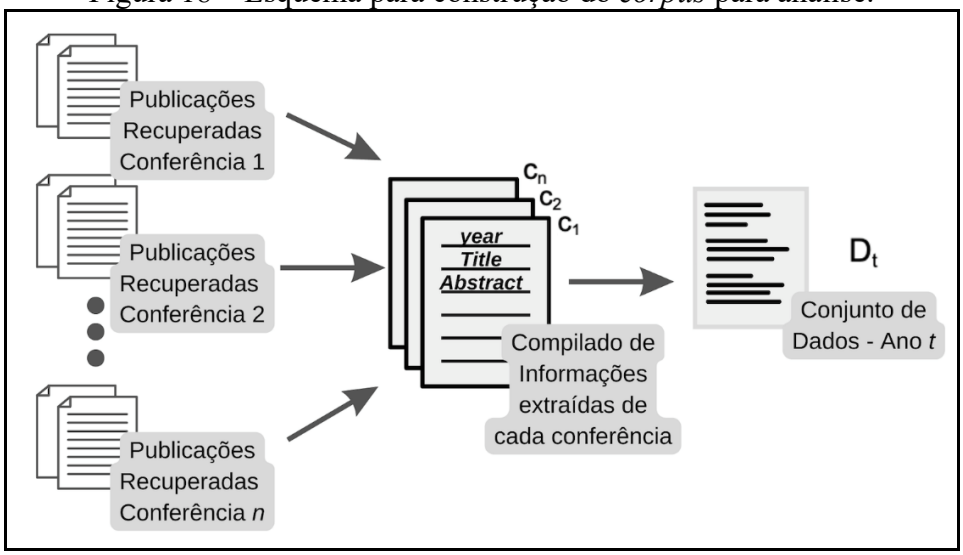
Figura 17 – Representação dos documentos recuperados e das informações extraídas.



Fonte: Elaborado pela autora

A Figura 18 esquematiza o processo de recuperação e extração da informação. Para cada ano tem-se um conjunto de dados $D_t = [c_1, c_2, ..., c_n]$, onde c_i, $1 \leq i \leq n$, é o compilado de informações extraídas de diferentes conferências *i* referente a um mesmo ano *t*. Em outras palavras, cada documento c_i é similar ao esquema da Figura 17 e o conjunto de dados final é a compilação dessas informações recuperadas e extraídas.

Figura 18 – Esquema para construção do *corpus* para análise.



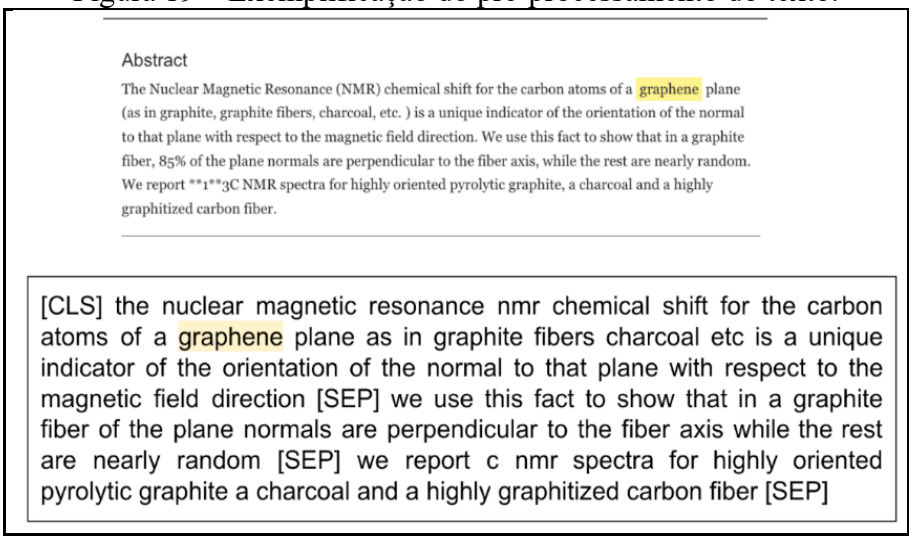
Fonte: Elaborado pela autora

(1.2) Pré-processamento

Tendo em vista o emprego do modelo de linguagem pré-treinado BERT, poucas tarefas de pré-processamento são adotadas, pois, o uso de tarefas, como *stemming* ou eliminação de *stopwords*, podem acarretar perda de informações importantes que contribuem para o aprendizado da rede neural. Por conseguinte, o arquivo de texto é o mais próximo possível do texto original. Além disso, o conjunto de dados do modelo constitui-se de publicações científicas, o que implica em uma linguagem formal e de maior confiança sem a necessidade de tratamentos adicionais nos dados.

São sugeridas a aplicação de funções de limpeza e transformação como: (i) uniformização do texto em letras minúsculas para uso do modelo BERT *base uncased*, (ii) remoção de caracteres especiais e algarismos e (iii) preparações de *preset* específicas para o modelo de linguagem escolhido. No caso, o modelo BERT necessita da divisão do texto em sentenças com *strings* de até 256 *tokens* e a inserção dos *tokens* especiais ([CLS] e [SEP]) que marcam o início e a separação entre sentenças. A Figura 19 apresenta um exemplo de pré-processamento do texto.

Figura 19 – Exemplificação do pré-processamento do texto.



Fonte: Elaborado pela autora

Essas escolhas de pré-processamento se justificam, pois a diferenciação da fonte não agrega nenhum tipo de significado para o presente problema. Por exemplo, novos acrônimos científicos que costumam ser grafados em maiúscula, quando difundidos, passam a ser grafados em minúscula, como no caso de *laser* ou *radar* (Seção 2.5), assim sendo, a falta de distinção de fonte não é considerada importante. A segunda, caracteres especiais e algarismos não agregam informações pertinentes para o propósito do modelo e seriam motivo de ruído no processamento. A terceira diz respeito à demanda particular da técnica de *embedding* escolhida.

(1.3) Disponibilizar *Corpus*

Como saída tem-se um conjunto de dados caracterizado pelos textos extraídos das publicações em conferências, referente ao ano de 1985, adequadamente pré-processados conforme as necessidades do modelo BERT, abordagem de NLP escolhida.

4.3.2 Etapa 2 - Extrair e Representar Palavras do Conjunto de Dados

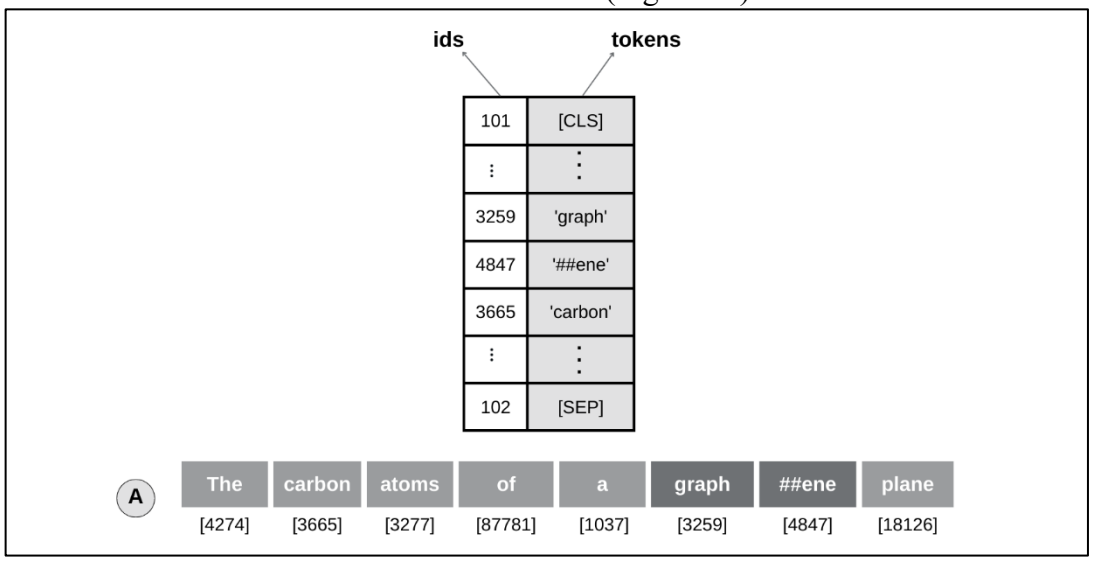
Com o conjunto de textos prontos torna-se necessário representar quantitativamente as palavras. Para a representação das palavras por *word embedding*, como já mencionado, optou-se pela implementação do modelo de linguagem BERT.

De início, o modelo BERT internamente executa a *token*ização do texto com a implementação do algoritmo *WordPieces* (WU *et al.*, 2016). Esse processo é responsável por particionar o texto em unidades de informação (*tokens*), relativos a palavras, subpalavras ou

caracteres, que o modelo é capaz de compreender. A *token*ização ocorre tal que, primeiro é verificado se a palavra inteira consta no vocabulário do BERT. Caso contrário, a palavra é particionada nas maiores subpalavras possíveis contidas no vocabulário.

Além da *token*ização, para seu processamento, o modelo BERT trabalha com os *ids* (identificadores) de cada palavra *token*izada, isto é, uma sequência de números inteiros que identifica cada *token* com seu número de índice no vocabulário de *tokens* do modelo (o vocabulário contém aproximadamente 30.000 *tokens* de palavras e subpalavras mais comuns encontradas no idioma inglês). Cabe notar que a associação de um número de identificação com um determinado *token* é estabelecido durante o processo de criação do vocabulário do modelo, atribuídos com base na frequência de ocorrência de cada *token* nos dados de treinamento (quanto maior a frequência de ocorrência, menor seu número de *id*), não exprimindo qualquer significado inerente ou valor semântico do *token*. Na Figura 20, o trecho A extraído do resumo (Figura 19), exemplifica um trecho do texto *token*izado e seus *ids*.

Figura 20 – Representação da associação de *ids* e *tokens* e a *token*ização de um trecho extraído do resumo (Figura 19).

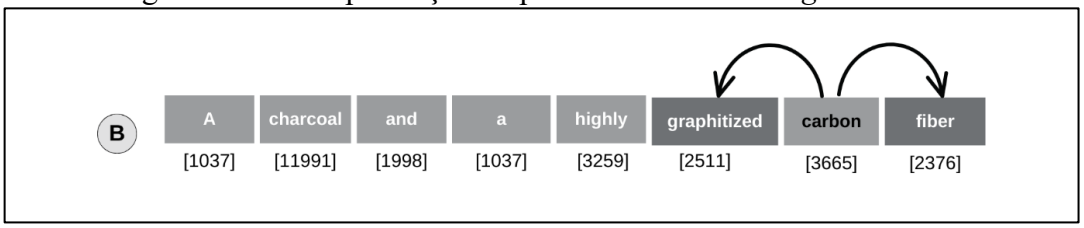


Fonte: Elaborado pela autora

Até o momento, para representação vetorial semântica das palavras, o texto foi apenas preparado (*token*izado) e identificado segundo o vocabulário do modelo de linguagem pré-treinado. Na sequência, o modelo é capaz de obter os *embeddings* contextuais das palavras, ajustando as 768 posições dos vetores de representação associados aos *tokens* durante a execução das tarefas padrões do modelo BERT (ML e NSP) com o conjunto de dados específico, pré-determinado e obtido na etapa anterior. A seguir, apresenta-se o racional embutido no processamento do modelo BERT.

Internamente, de modo singelo, para obter o *embedding* contextual, considerando o trecho (B) (Figura 21), a palavra '*carbon*' é a palavra alvo para obter o *embedding* e os pares {'*carbon'*, '*graphitized'*} e {'*carbon'*, '*fiber'*} são os termos que compõem seu contexto.

Figura 21 – Exemplificação do processo de *embedding* contextual.

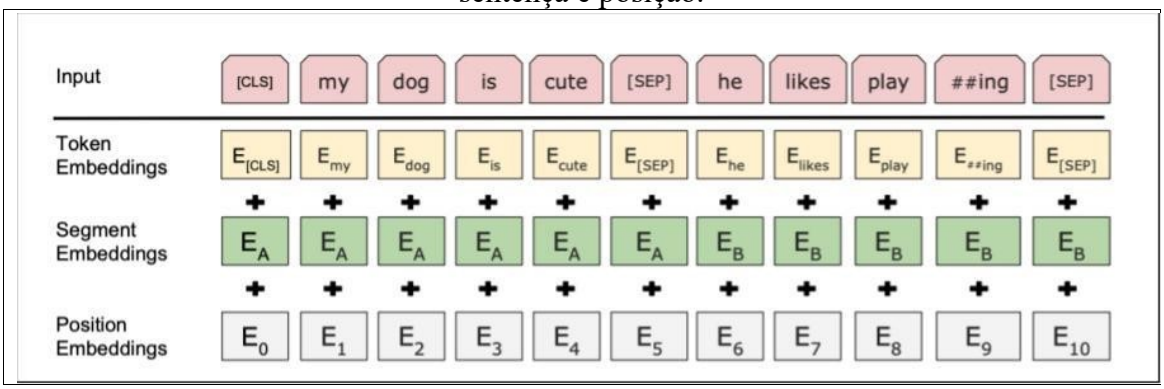


Fonte: Elaborado pela autora

O texto não é uma sequência aleatória de palavras, ou seja, a ordem de emprego das palavras é o que permite a compreensão do sentido. Sendo assim, é esta capacidade de capturar o relacionamento e o contexto em que as palavras ocorrem, que faz da representação vetorial ofertada pelo modelo BERT que sejam *embeddings* semânticos contextuais dinâmicos. Para realizar esse "aprendizado", o modelo BERT sabe a posição em que as palavras aparecem e executa as tarefas de prever tanto a palavra, quanto à sentença seguinte (seção 2.3.1.1). Nesse sentido, a representação de cada *token* é um agrupamento de outros três *embeddings* calculados internamente pelo modelo (Figura 22):

- *Token embeddings* - relativo aos *ids* de vocabulário e a associação de um vetor de *embedding* ao *token* por meio de uma matriz *token-embeddings* aprendida;
- *Segment embeddings* - diferenciam quais palavras são relativas à primeira sentença e quais são associadas à sentença subsequente; e
- *Position embeddings* - indicam a posição relativa de cada *token* na sequência.

Figura 22 – O *embedding* do modelo BERT é obtido a partir dos *embeddings* de *token*, sentença e posição.



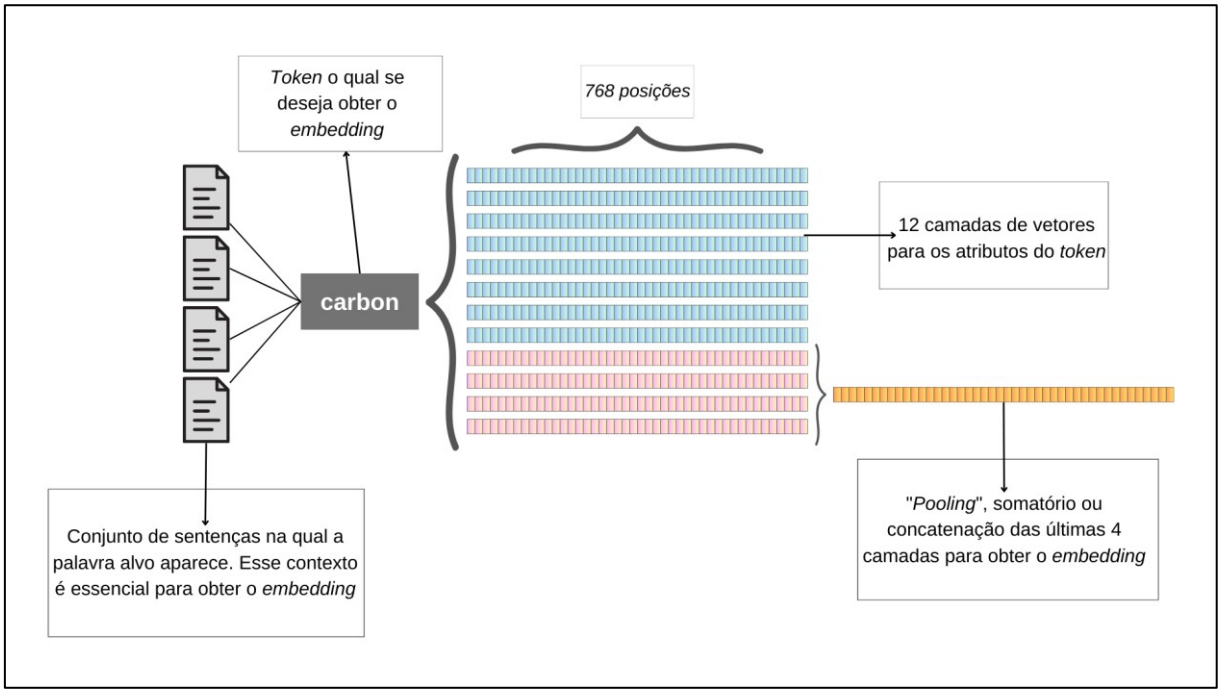
Fonte: Extraído de Devlin *et al*. (2019)

Convém observar que, palavras que não fazem parte do vocabulário são representadas como subpalavras marcadas pelo caractere '##' (dupla cerquilha), como a palavra '*graphene*' no trecho A (Figura 20). Em consequência disso, existe uma discussão sobre como obter o *embedding* da palavra e não apenas de suas subpartes, essa questão é retomada na etapa 3 subsequente (seção 4.3.3). Neste momento, concentra-se em elucidar como ocorrem os *embeddings* de palavras que existem no vocabulário, como no trecho B (Figura 21).

O texto *token*izado de entrada é então processado através de 12 camadas de *'transformers'* (modelo BERT *base*) obtendo uma representação multicamada dos *tokens*. Essa etapa de codificação (*encoding*) é um processo complexo de *deep learning,* incluindo camadas de *'self-attention'* e *'feed-forward'*. Em síntese, cada camada *'transformer'* é responsável por realizar uma série de operações que refinam a representação de cada *token* na sequência para capturar e exprimir as relações contextuais.

Então, para cada *token*, o modelo tem como saída 12 camadas de vetores em que, para obter o *embedding* é necessário combiná-las. Essa estratégia é conhecida como *pooling,* sendo sugerido por Devlin *et al*. (2019) concatenar ou somar as 4 últimas camadas para obter os *embeddings,* pois são consideradas as camadas que contêm as informações mais relevantes para a tarefa de extração de atributos (Figura 23).

Figura 23 – Esquema do *embedding* de palavras conhecidas pelo vocabulário do modelo BERT.



Fonte: Elaborado pela autora

Como resultado desta etapa tem-se a representação das palavras ou sub-palavras presentes no texto, identificadas como *tokens* presentes no vocabulário do modelo BERT, em vetores numéricos capazes de absorver o contexto de ocorrência das palavras.

4.3.3 Etapa 3 - Identificar Sinais Novos

A identificação de sinais novos implica na descoberta de novas palavras. Considerando os modelos de linguagem pré-treinados como o BERT, se mostra um desafio identificar, inserir e obter o *embedding* dessas palavras desconhecidas. Frizando que nesta demostração compreende por palavra nova apenas a palavra inédita.

Modelos de linguagem pré-treinados, estabelecidos em contextos gerais, apresentam baixo desempenho em domínios específicos (HU *et al*., 2019). Com base no argumento que se o modelo não foi pré-treinado em um conjunto de textos que apresentava a palavra em uma frequência relevante, o resultado é a geração de vetores de *embedding* imprecisos.

O mesmo ocorre quando o modelo se depara com palavras OOV ou raras. No caso, o modelo BERT irá particionar a palavra desconhecida em subpalavras conhecidas pelo seu vocabulário. Cada *token* de subpalavra é processado separadamente e a representação final da palavra pode ser obtida pela média das representações de todos seus *tokens* de subpalavra. No entanto, embora o modelo BERT seja capaz de gerar um *embedding* para uma palavra desconhecida, a qualidade dessa representação não apresenta a mesma qualidade das palavras reconhecidas pelo vocabulário. De modo geral, a soma dos *embeddings* que exprimem o contexto de ocorrência das subpalavras não é capaz de capturar o significado semântico da palavra unificada, fator importante para o modelo considerando a análise temporal das palavras.

Muitas alternativas para lidar com essas questões de palavras OOV ou raras baseiam-se em contornar o problema, por exemplo, trocando uma palavra desconhecida por um sinônimo, como o algoritmo PatchBERT (MOON; OKAZAKI, 2020). Mas, essa não é uma opção quando o intuito é aprender novas palavras, como a proposta dessa pesquisa, identificando novos termos do domínio de conhecimentos científicos e tecnológicos.

Na literatura pode-se encontrar também algoritmos que visam ampliar os *embeddings* do modelo BERT pré-treinado-o em domínios específicos, como o BioBERT® (LEE *et al.*, 2020) e o SciBERT® (BELTAGY; LO; COHAN, 2019), para que assim sejam compreendidas

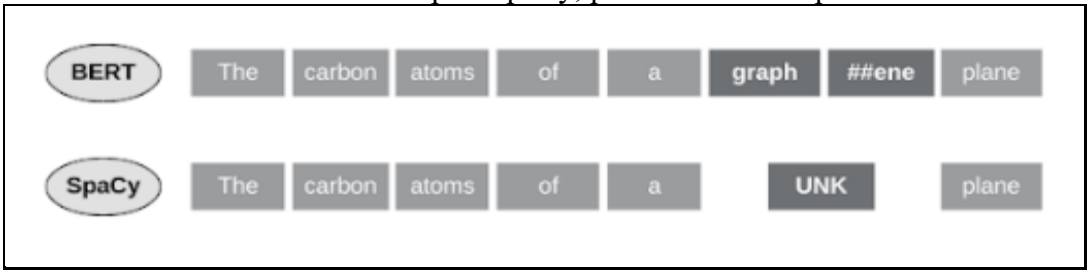
como um único *token* palavras que no modelo BERT original seriam particionadas em *tokens* de subpalavras. Esta abordagem resolve o problema de *embeddings* imprecisos para domínios específicos, mas não lidam com o problema de identificar e incluir novas palavras, pois ainda só aprendem palavras com certa frequência no *corpus* de treinamento. Mais próximo a esse desafio encontrou-se os algoritmos BERTRAM (SCHICK; SCHÜTZE, 2020) e, especialmente, o exBERT (TAI *et al.*, 2020), que propõe uma alteração na arquitetura original do modelo BERT, utilizando módulos de extensão para incorporar o vocabulário do domínio específico reutilizando parâmetros do modelo original pré-treinado, possibilitando o *embedding* de novas palavras, e não apenas das subpalavras.

(3.1) Identificar Palavras Novas

Para a descoberta das palavras recorreu-se a uma combinação do *token*izador spaCy® e do modelo BERT (DEVLIN *et al.*, 2019). Com a escolha do modelo BERT para obter os *embeddings* tem-se que uma palavra pode ser composta por mais de um *token*. No exemplo, o trecho (A) (Figura 20) apresenta a palavra '*graphene*' como desconhecida pelo vocabulário do modelo, portanto, o *token*izador *WordPiece* a particiona em *tokens* de subpalavras conhecidas: '*graph*' e '*##ene*', marcada pela presença do símbolo '##'. Em contrapartida, o *token*izador spaCy não trabalha com *token*ização de subpalavras, dessa maneira, ao se deparar com uma palavra desconhecida ela será rotulada como 'UNK' (desconhecida).

Dessa maneira, ao comparar um mesmo trecho *token*izado por cada uma das abordagens, pode-se descobrir quais palavras são novas. Por exemplo, a *token*ização do trecho (A) "*[...] carbon atoms of a graphene plane [...]*", no spaCy fica ['carbon', 'atoms', 'of, 'a', 'UNK', 'plane'] e ao comparar com a *token*ização oferecida pelo modelo BERT ['carbon', 'atoms', 'of, 'a*, 'graph' '##ene', 'plane']*, identifica-se que a palavra desconhecida [UNK] é composta pelas subpalavras [graph] e [##ene], como pode ser visto na Figura 24.

Figura 24 – Representação da diferença no *embedding* de um mesmo trecho pelos *token*izadores do modelo BERT e pelo spaCy, para identificar a palavra desconhecida.



Fonte: Elaborado pela autora

Uma vez identificada a palavra nova, para obter seu *embedding* sugere-se a abordagem oferecida pelo exBERT de Tai *et al*. (2020). O algoritmo opera combinando a saída dos módulos originais BERT e de extensão exBERT, "aprendendo" novas palavras para não serem partidas em *tokens* de subpalavras em momentos subsequentes.

(3.2) Classificar como 'sinal novo'

Como saída desta etapa são identificadas as palavras novas e seus *embeddings*. Neste momento, o modelo identifica a palavra '*graphene*' como uma palavra nova e a armazena em uma lista, organizada por ano, classificando-a como um 'sinal novo'. Essa lista é importante para a etapa de monitoramento.

4.3.4 Etapa 4 - Elaborar Rede de Palavras

Esta etapa objetiva conceber a rede formada pelas palavras que compõem o conjunto de dados, evidenciando como estas se relacionam entre si. Os *embeddings* das palavras implicam o posicionamento das mesmas em um espaço vetorial e a medida de similaridade estabelece a intensidade da relação entre as palavras.

(4.1) Calcular a Similaridade

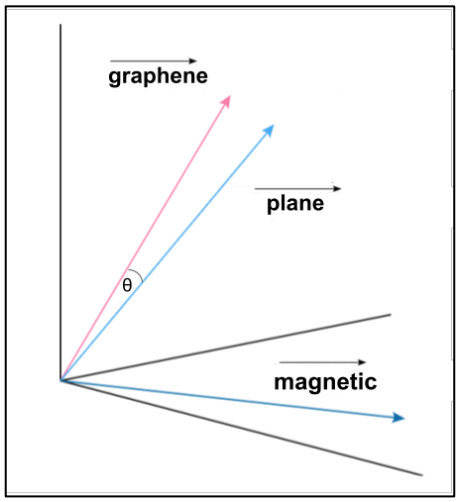
Existem diferentes medidas para calcular a distância entre vetores: Euclidiana, Cosseno, *Jaccard*, *Mahalanobis*, *Manhattan e Hamming* são alguns exemplos. Contudo, a medida de similaridade por cossenos é a mais implementada nesses casos, em virtude de ser mais informativa para vetores de alta dimensionalidade, como os *embeddings* de palavras do modelo BERT.

O cálculo de similaridade pelo cosseno (Equação 4) baseia-se na distância angular entre dois vetores e retorna valores entre -1 e 1, onde 1 corresponde à similaridade perfeita e -1 dissimilaridade total. É importante ressaltar que uma alta similaridade não implica em dois termos serem parecidos, mas em estarem relacionados por compartilharem aspectos em comum. São mais próximos em contraste com um terceiro elemento, em outras palavras, a medida não é absoluta, mas relativa. Não diz respeito se A é parecido com B, mas que A é mais próximo a B do que C, por exemplo.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

Considerando o trecho (A) do exemplo anterior (Figura 20), o par {‘*graphene’, ‘plane’*} terá uma similaridade maior, pois ocorrem com mais probabilidade em um mesmo contexto do que as palavras {‘*graphene’, ‘magnetic’*}. Um alto valor de medida de similaridade por cossenos não é sobre dois termos significarem a mesma coisa, mas por pertencerem a um mesmo contexto em um espaço *n*-dimensional. A Figura 25 ilustra como o vetor da palavra *‘graphene’* é mais próximo da palavra *‘plane’*, ou seja, maior similaridade se comparado ao vetor da palavra *‘magnetic’*.

Figura 25 – Proximidade entre os vetores de palavras para o cálculo de similaridade pela medida do cosseno.

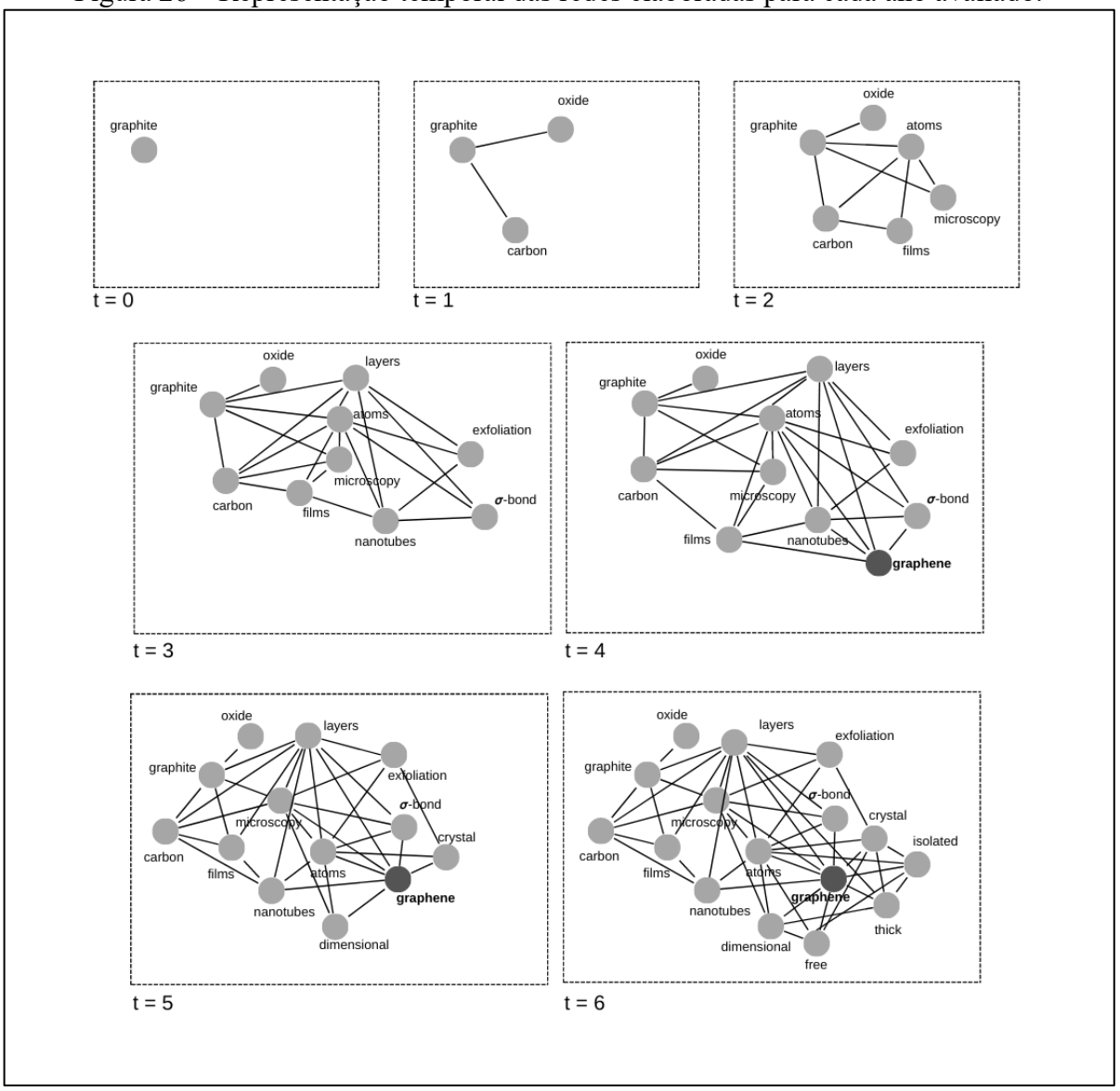


Fonte: Elaborado pela autora

(4.2) Exibir a Rede de Palavras

Na construção da rede de palavras, os nós são os vetores de *embedding* das palavras e as arestas de conexão a similaridade entre elas. Tomando como base que contextos semelhantes têm representações vetoriais semelhantes, de tal forma que, palavras que ocorrem com mais probabilidade em um mesmo contexto são mais próximas, estabelece-se uma medida de corte, tal que as arestas são representadas conectando dois nós apenas se a similaridade for acima desse limiar. Dessa forma, palavras com baixa similaridade não aparecerão relacionadas (conectadas) na rede, sendo representadas, portanto, apenas as palavras mais próximas. A Figura 26 demonstra como seria uma sequência temporal de redes de palavras ilustrando o surgimento da palavra ‘*graphene*’, no momento $t = 4$.

Figura 26 – Representação temporal das redes elaboradas para cada ano avaliado.



Fonte: Elaborado pela autora

4.3.5 Etapa 5 - Identificar Sinais Latentes

Essa etapa tem o propósito de monitorar a rede de palavras, identificando um crescimento significativo nas relações em torno das palavras novas como reflexo do interesse no contexto no qual estas se inserem, no intuito de distinguir os sinais novos que apresentam crescimento (sinais latentes), dos sinais novos que podem ser considerados apenas ruídos.

(5.1) Monitorar a Dinamicidade Temporal da Rede

Para monitoramento, o modelo incorpora algoritmos de análise de redes temporais como o modelo *Top-k Density Burst Subgraphs Finding* (TDF) de Chu *et al*. (2019) ou de Li *et al.* (2022) "*Early Bursting Cohesive Subgraph Discover*y", permitindo a identificação de

subgrafos que apresentam uma densidade acumulada rapidamente em um determinado intervalo de tempo.

Para o exemplo *'graphene',* conceitualmente não é possível implementar tais técnicas, mas pode-se observar o crescimento das publicações como um indício de crescimento de interesse que seria refletido no volume de nós e conexões na rede. A Tabela 1 foi elaborada a partir das publicações em congresso que apresentam a palavra '*graphene*' em seu título ou resumo, obtidas na base *Scopus*. Tomando como base o crescimento no volume de menções da palavra '*graphene*', pode-se supor que o comportamento *burst* seria detectado no ano de 2001 ou 2002. Desta forma, relacionando a Figura 26 com a Tabela 1, $t = 4$ seria 1985, $t = 5$ seria 1995/1996, onde, verifica-se um pequeno crescimento nas publicações e $t = 6$ o ano de 2001/2002 quando há um crescimento expressivo no volume de publicações.

Tabela 1 – Volume de publicações contendo a palavra '*graphene*' na base *Scopus*.

Ano	Número de publicações	Ano	Número de publicações	Ano	Número de publicações
1985	1	1996	5	2001	15
1991	3	1997	0	2002	24
1993	1	1998	4	2003	22
1994	1	1999	2	2004	39
1995	5	2000	1	2005	61

Fonte: Elaborado pela autora

(5.2) Classificar como 'sinal latente'

Como pode ser visto na imagem do exemplo (Figura 26), a rede em torno do termo '*graphite*' cresce ao longo do tempo, se tornando densa a uma certa velocidade (*burst*), e inclui um sinal novo em seu subgrafo, o '*graphene*'. Em $t = 6$ o termo '*graphene*' passa, então, a ser não somente um sinal novo, mas também um sinal latente. Neste sentido, a palavra é removida da lista de sinais novos e incluída na lista de sinais latentes.

4.3.6 Etapa 6 - Identificar Sinais Fracos

Esta etapa visa estimar, dentre os sinais latentes, aqueles que apresentam um potencial impacto tecnológico. Desse modo, a etapa inclui a recuperação e preparação dos documentos de pedidos de patentes, dos quais são extraídas as palavras a fim de elaborar uma lista de palavras tecnológicas.

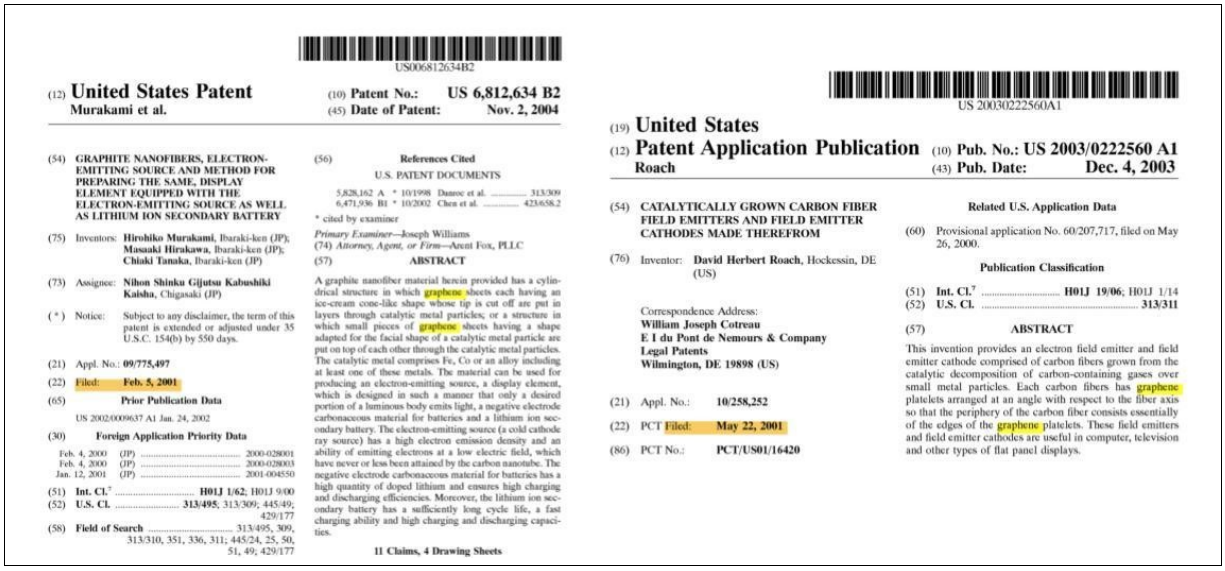
(6.1) Estimar Potencial Impacto dos Sinais Latentes

A informação tecnológica pode ser obtida de diferentes bases. Segundo Singh, Chakraborty e Vincent (2016), os principais repositórios para depósito de patente são: *United States Patent and Trademark Office* (USPTO®), *European Patent Office* (EPO®) e *Japan Patent Office* (JPO®). Contudo, a base USPTO é usualmente eleita, considerando que reivindicações enviadas para escritórios de um determinado país são frequentemente enviadas simultaneamente aos Estados Unidos, se mostrando uma base expressiva para o mercado tecnológico ao nível internacional (BASS; KURGAN, 2010).

Logo, são recuperados os documentos de pedido de patente do respectivo ano do qual também foram recuperadas as publicações de congressos. Por exemplo, como mencionado no início desta demonstração, considerando janeiro de 1986, as publicações de congressos são referentes ao ano anterior de 1985. Idem para as publicações de patentes, que no início de 1986 seriam extraídas informações dos pedidos de patente registrados no ano de 1985 e adicionadas a essa lista de "palavras tecnológicas". A Figura 27 apresenta os dois documentos de patentes que mencionam pela primeira vez a palavra '*graphene*' em seu resumo (US006812634B2 e US20030222560A1). Lembrando que foi considerada a data do pedido de patente, não o ano em que foi concedida.

Na sequência, com a lista de palavras tecnológicas, é realizada uma consulta a partir da lista de sinais latentes. Acompanhando o raciocínio da etapa anterior, na qual, a palavra '*graphene*' seria identificada como um sinal latente em 2001 ou 2002, concomitante é verificada sua primeira menção em um resumo de pedido de patente na base USPTO.

Figura 27 – Primeiros documentos de pedido de patentes que apresentam o termo ‘*graphene*’ em seus resumos.



US006812634B2

(12) **United States Patent**
Murakami et al.

(10) **Patent No.: US 6,812,634 B2**
(45) **Date of Patent: Nov. 2, 2004**

(54) **GRAPHITE NANOFIBERS, ELECTRON-EMITTING SOURCE AND METHOD FOR PREPARING THE SAME, DISPLAY ELEMENT EQUIPPED WITH THE ELECTRON-EMITTING SOURCE AS WELL AS LITHIUM ION SECONDARY BATTERY**

(75) Inventors: **Hirohiko Murakami**, Ibaraki-ken (JP); **Masaaki Hirakawa**, Ibaraki-ken (JP); **Chiaki Tanaka**, Ibaraki-ken (JP)

(73) Assignee: **Nihon Shinku Gijutsu Kabushiki Kaisha**, Chigasaki (JP)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 550 days.

(21) Appl. No.: **09/775,497**

(22) Filed: **Feb. 5, 2001**

(65) **Prior Publication Data**

US 2002/0009637 A1 Jan. 24, 2002

(30) **Foreign Application Priority Data**

Feb. 4, 2000 (JP) 2000-028001
Feb. 4, 2000 (JP) 2000-028003
Jan. 12, 2001 (JP) 2001-004550

(51) **Int. Cl.**[7] **H01J 1/62**; H01J 9/00
(52) **U.S. Cl.** **313/495**; 313/309; 445/49; 429/177
(58) **Field of Search** 313/495, 309, 313/310, 351, 336, 311; 445/24, 25, 50, 51, 49; 429/177

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,828,162 A * 10/1998 Daaroc et al. 313/309
6,471,936 B1 * 10/2002 Chen et al. 423/658.2

* cited by examiner

Primary Examiner—Joseph Williams
(74) *Attorney, Agent, or Firm*—Arent Fox, PLLC

(57) **ABSTRACT**

A graphite nanofiber material herein provided has a cylindrical structure in which graphene sheets each having an ice-cream cone-like shape whose tip is cut off are put in layers through catalytic metal particles; or a structure in which small pieces of graphene sheets having a shape adapted for the facial shape of a catalytic metal particle are put on top of each other through the catalytic metal particles. The catalytic metal comprises Fe, Co or an alloy including at least one of these metals. The material can be used for producing an electron-emitting source, a display element, which is designed in such a manner that only a desired portion of a luminous body emits light, a negative electrode carbonaceous material for batteries and a lithium ion secondary battery. The electron-emitting source (a cold cathode ray source) has a high electron emission density and an ability of emitting electrons at a low electric field, which have never or less been attained by the carbon nanotube. The negative electrode carbonaceous material for batteries has a high quantity of doped lithium and ensures high charging and discharging efficiencies. Moreover, the lithium ion secondary battery has a sufficiently long cycle life, a fast charging ability and high charging and discharging capacities.

11 Claims, 4 Drawing Sheets

US 20030222560A1

(19) **United States**
(12) **Patent Application Publication**
Roach

(10) **Pub. No.: US 2003/0222560 A1**
(43) **Pub. Date: Dec. 4, 2003**

(54) **CATALYTICALLY GROWN CARBON FIBER FIELD EMITTERS AND FIELD EMITTER CATHODES MADE THEREFROM**

(76) Inventor: **David Herbert Roach**, Hockessin, DE (US)

Correspondence Address:
William Joseph Cotreau
E I du Pont de Nemours & Company
Legal Patents
Wilmington, DE 19898 (US)

(21) Appl. No.: **10/258,252**

(22) PCT Filed: **May 22, 2001**

(86) PCT No.: **PCT/US01/16420**

Related U.S. Application Data

(60) Provisional application No. 60/207,717, filed on May 26, 2000.

Publication Classification

(51) **Int. Cl.**[7] **H01J 19/06**; H01J 1/14
(52) **U.S. Cl.** **313/311**

(57) **ABSTRACT**

This invention provides an electron field emitter and field emitter cathode comprised of carbon fibers grown from the catalytic decomposition of carbon-containing gases over small metal particles. Each carbon fibers has graphene platelets arranged at an angle with respect to the fiber axis so that the periphery of the carbon fiber consists essentially of the edges of the graphene platelets. These field emitters and field emitter cathodes are useful in computer, television and other types of flat panel displays.

Fonte: Recuperado a partir do *Google Patents*®

(6.2) Classificar como ‘sinal fraco’

Quando a palavra ‘*graphene*’, presente na lista de sinais latentes, é encontrada na lista de palavras extraídas dos pedidos de patentes, ela passa a ser considerada um sinal fraco e deixa a lista de sinais latentes. Portanto, em 2001 ou 2002 a palavra ‘*graphene*’ seria sugerida como um sinal fraco para tecnologias emergentes.

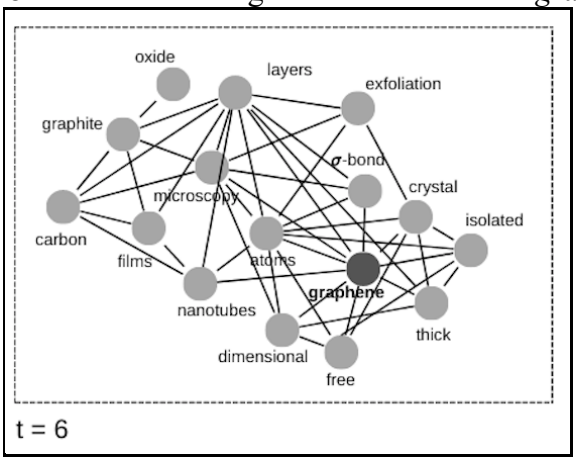
Convém mencionar como ressalva desta etapa, que embora apontado que informações provenientes de relatórios de consultorias ou notícias de websites seriam interessantes fontes complementares de conhecimento tecnológico, esse conhecimento não é de fácil acesso para uma demonstração em retrospectiva. Em função disso, tomamos os pedidos de patentes como marcos para a exemplificação. No entanto, em uma aplicação no presente visando o futuro seria pertinente a inclusão de tais informações.

4.3.7 Etapa 7 - Apresentar Recomendações de Sinais Fracos para Tecnologias Emergentes

Como entrega dos sinais fracos detectados, sugere-se a visualização do domínio tecnológico que essa palavra nova pertence. Em outras palavras, recuperar o subgrafo ao qual pertence o sinal fraco identificado para visualizar o contexto que o descreve visando identificar a possível tecnologia emergente, como evento futuro revelado pelo sinal fraco (Figura 28). Ou ainda, sugere-se visualizar a trajetória de evolução temporal dos sinais.

Nesse momento, observaria-se que a palavra nova identificada "*graphene*" remete a um contexto de finos cristais de grafite. Promovendo indícios de que talvez seja relevante buscar mais informações sobre esse nicho tecnológico, suas aplicações e implicações.

Figura 28 – Recorte do subgrafo do sinal latente '*graphene*'.

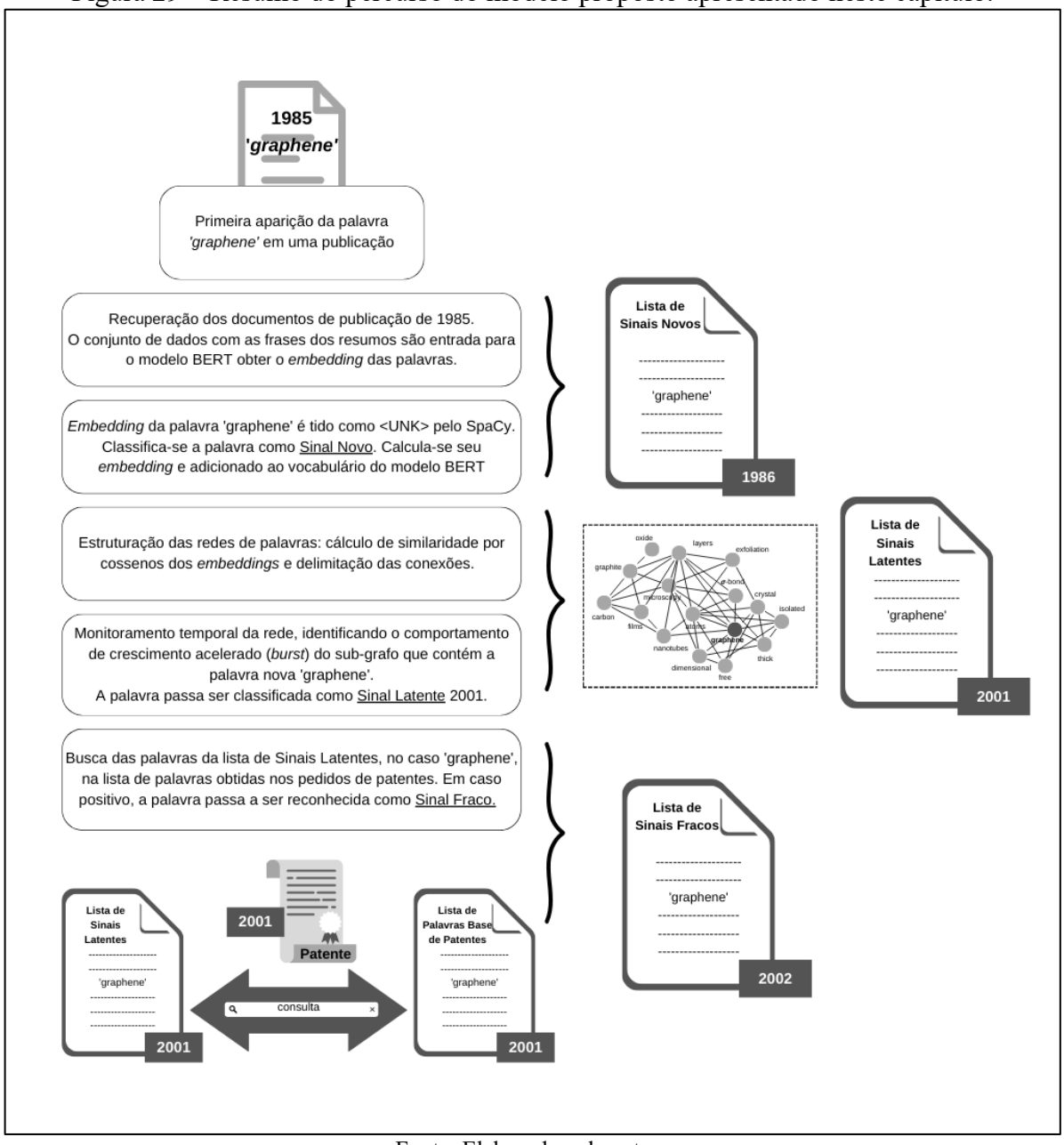


Fonte: Elaborado pela autora

Concluindo, a palavra '*graphene*' seria identificada como um sinal novo em 1985; monitorada até o ano de 2001, 2002, quando seria detectado o crescimento (*burst*) no subgrafo ao qual pertence, sendo classificada como um sinal latente. E, por fim, em 2001 ou 2002, seria classificada como um sinal fraco devido a sua aparição na base de patentes. Constata-se ainda que, embora o modelo não identifique a palavra '*graphene*' como um sinal fraco próximo à data considerada de sua primeira patente (1991), o momento no qual, supostamente, o modelo sugere '*graphene*' como um sinal fraco para tecnologias emergentes (2001) se mostra notável. Em consonância com as informações históricas levantadas na introdução da seção, e observando o gráfico de publicações (Figura 16), o sinal fraco seria sugerido quase uma década antes do início do crescimento exponencial nas pesquisas, após o Prêmio Nobel de 2010.

Para finalizar, um esquema com o fluxo do modelo é apresentado na Figura 29. Do lado esquerdo está o ano e o fluxo de tarefas, do lado direito as principais saídas das etapas e o ano correspondente em que as informações seriam descobertas. Apresenta-se, também, uma síntese no Quadro 15 para as técnicas e ferramentas adotadas ou sugeridas em cada etapa da demonstração. Vale lembrar que o modelo é iterativo e transpassa por todas as etapas ano a ano, de modo que o conjunto de dados analisado corresponde aos documentos recuperados a cada ano. Aqui, a título de exemplo, realizou-se uma análise retrospectiva na intenção de verificar a capacidade do modelo em atender seu objetivo. Entretanto, o modelo foi elaborado para monitorar as informações do presente em direção ao futuro.

Figura 29 – Resumo do percurso do modelo proposto apresentado neste capítulo.



Fonte: Elaborado pela autora

Quadro 15 – Síntese das técnicas adotadas para execução do modelo proposto.

	Etapa	Técnicas e Ferramentas
1	Elaborar o Conjunto de Dados	Ferramentas como o Wget®, DAP®, NLTK®, MongoDB®
2	Extrair e Representar Palavras do Conjunto de Dados	BERT® (DEVLIN *et al.*, 2019)
3	Identificar Sinais Novos	SpaCy® e ExBERT® (TAI *et al.*, 2020)
4	Elaborar a Rede de Palavras	Similaridade por cossenos
5	Identificar Sinais Latentes	Análise temporal *Burst* de subgrafos - Algoritmo TDF (*Top-k Density Burst Subgraphs Finding)* (CHU *et al.*, 2019)
6	Identificar Sinais Fracos	Consulta na base de patentes
7	Apresentar as Recomendações de Sinais Fracos para Tecnologias Emergentes	Softwares ou bibliotecas de visualização de redes e visualização temporal como o Gephi®.

Fonte: Elaborado pela autora

5 AVALIAÇÃO EXPERIMENTAL DO MODELO

Como declarado no Capítulo 3, a DSRM conta com uma etapa de avaliação que pode recorrer a diferentes meios para avaliar os artefatos propostos. Neste trabalho, adotou-se uma avaliação descritiva e uma experimental. A primeira foi apresentada junto ao modelo, na seção 4.2, valendo-se das justificativas teóricas para proposição das etapas que compõem o modelo.

A avaliação experimental apresentada neste capítulo visa apreciar o modelo enquanto sua capacidade de solução em atenter o problema levantado. Nesse sentido, buscou-se mostrar a operacionalização do encadeamento ilustrado conceitualmente na seção anterior de demonstração. Como cenário de estudo, simulou-se o modelo considerando a tecnologia VoIP em retrospectiva. A palavra *'voip'* foi eleita como 'palavra nova' para estudo por se tratar de uma tecnologia relativamente recente e que se estabeleceu ao decorrer dos anos. Ademais, a palavra tornou-se apta para o estudo uma vez que não está nativamente inclusa no vocabulário BERT.

Considera-se pertinente mencionar que envolvem processos de decisão diferentes para o desenvolvimento do projeto do produto (modelo) e decisões para o projeto de protótipo (operacionalização). Ao passo que as decisões de elaboração do modelo foram apresentadas na seção 4.2.2, e esta seção apresenta decisões de protótipo para viabilizar a implementação computacional do modelo. Isto posto, inicialmente contextualiza-se historicamente a tecnologia VoIP e, em seguida, são descritas as etapas do modelo exemplificando sua operacionalização.

5.1 TECNOLOGIA VOIP

A tecnologia VoIP, do inglês ***V**oice **o**ver **I**nternet **P**rotocol* (em português traduzido como, **Vo**z sobre **IP**), permite realizar chamadas de voz usando uma conexão de *Internet* de banda larga ao invés de uma linha telefônica normal (ou analógica). Concisamente, a técnica VoIP converte a voz em um sinal digital que viaja pela *Internet*, em outras palavras, telefonia em nuvem. Isso quer dizer uma telefonia que se apropria da estrutura da *Internet* para fornecer seus serviços, chamadas de voz e vídeo (POURGHASEM; KARIMI; EDALATPANAH, 2012). Por conseguinte, são requisitos para o funcionamento da tecnologia uma conexão de banda larga (*Internet* de alta velocidade) e um computador com microfone acoplado ou telefone especializado, por exemplo, os *smartphones*.

A ideia de transferir informações através de pacotes baseados em IP foi atraente pela sua velocidade, melhora na qualidade e menor custo. Considera-se 1995 como o ano de início operacional do VoIP com a empresa VocalTec®, o aplicativo foi chamado de *Internet Phone*. No ano seguinte, foram criados os recursos de correio de voz pela *Internet*. Nesse momento da história do VoIP foi possível enviar mensagens de voz através da *Internet* para um telefone de destino. De início, eram comuns inúmeros problemas como qualidade devido a ruídos, períodos de silêncio ou atrasos (*delays*) e perda de conexão. Contudo, com os avanços tecnológicos (*softwares* de voz/vídeo chamada, *wi-fi, smartphones*, *tablets*, fibra ótica, aplicativos de mensagens instantâneas, etc), conforme o esperado, a tecnologia VoIP cresceu de forma rápida. Até o final de 1998, as chamadas VoIP representavam menos de 1% de todas as chamadas de voz. Em 2003, o número de chamadas VoIP cresceu para 25% de todas as chamadas de voz, somente nos EUA (SINGH *et al.*, 2014). Atualmente, chamadas via internet por meio de *smartphones* são recursos usados diariamente.

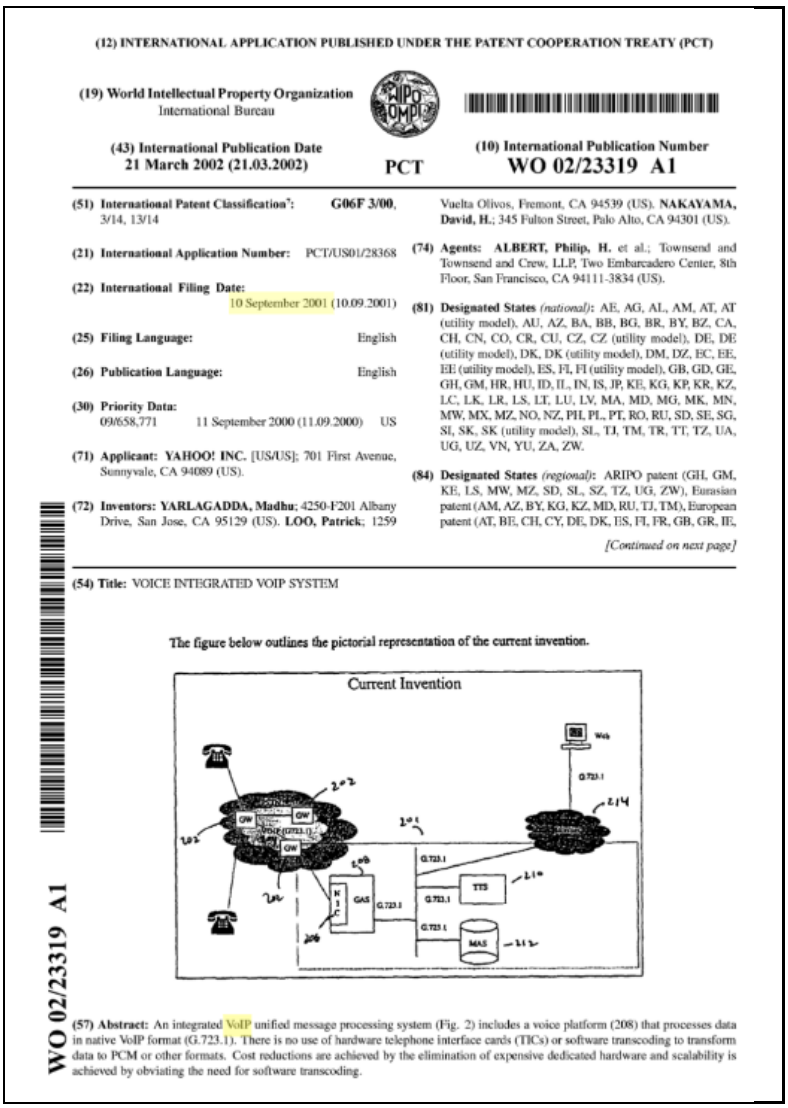
A título de curiosidade, pode-se considerar como marco inicial do VoIP o ano de 1974. Neste ano a *Advanced Research Projects Agency* (ARPANET), responsável pelo desenvolvimento precursor da *Internet*, enviou uma amostra de voz unidirecional em tempo real de 16 kb/s entre computadores localizados no *Information Sciences Institute* e no *Lincoln Lab*, EUA. Os algoritmos LPC (*Linear Predictive Coding*) implementados na época seguem sendo um padrão de codificação de áudio na tecnologia VoIP moderna.

Em 1974, o *Institute of Electrical and Electronics Engineers* (IEEE) publicou um artigo intitulado: "*A Protocol for Packet Network Interconnection*", marcando uma série de estudos posteriores voltados ao mesmo propósito, apesar de não empregar o acrônimo VoIP. Outro marco relevante se deu em janeiro de 2010, quando a Apple Inc.® atualizou o SDK (*Software Development Kit*) de desenvolvimento do iPhone® para permitir VoIP em redes celulares. O *iCall*® se tornou o primeiro aplicativo da *App Store*® a habilitar VoIP no *iPhone*® e *iPod Touch*® em redes 3G de celular. Desse modo, em quase 3 décadas a tecnologia surgiu e se disseminou, a ponto de já fazer parte do cotidiano de muitas pessoas, tal que a tecnologia, muitas vezes, passa despercebida.

Historicamente, a primeira patente da tecnologia VoIP é atribuída à Alon Cohen, co-fundador da VocalTec em 1995 (US5825771A). Entretanto, a patente apesar de se referir a criação de produtos de *Voice Over Networks*, eventualmente, originando a indústria de VoIP, ela não menciona o acrônimo. A partir de uma busca no *Google Patents* com a palavra '*voip*'

foi identificado um pedido de patente da Yahoo inc.® em 2001 (WO2002023319A1) como sendo a primeira patente a mencionar em seu resumo o acrônimo '*voip*' (Figura 30).

Figura 30 – Primeira página do pedido de patente que apresenta a palavra '*voip*' em seu resumo.



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 March 2002 (21.03.2002)

PCT

(10) International Publication Number
WO 02/23319 A1

(51) International Patent Classification⁷: G06F 3/00, 3/14, 13/14

(21) International Application Number: PCT/US01/28368

(22) International Filing Date: 10 September 2001 (10.09.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/658,771 11 September 2000 (11.09.2000) US

(71) Applicant: YAHOO! INC. [US/US]; 701 First Avenue, Sunnyvale, CA 94089 (US).

(72) Inventors: YARLAGADDA, Madhu; 4250-F201 Albany Drive, San Jose, CA 95129 (US). LOO, Patrick; 1259 Vuelta Olivos, Fremont, CA 94539 (US). NAKAYAMA, David, H.; 345 Fulton Street, Palo Alto, CA 94301 (US).

(74) Agents: ALBERT, Philip, H. et al.; Townsend and Townsend and Crew, LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111-3834 (US).

(81) Designated States (national): AE, AG, AL, AM, AT, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, CZ (utility model), DE, DE (utility model), DK, DK (utility model), DM, DZ, EC, EE, EE (utility model), ES, FI, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,

[Continued on next page]

(54) Title: VOICE INTEGRATED VOIP SYSTEM

The figure below outlines the pictorial representation of the current invention.

Current Invention

(57) Abstract: An integrated VoIP unified message processing system (Fig. 2) includes a voice platform (208) that processes data in native VoIP format (G.723.1). There is no use of hardware telephone interface cards (TICs) or software transcoding to transform data to PCM or other formats. Cost reductions are achieved by the elimination of expensive dedicated hardware and scalability is achieved by obviating the need for software transcoding.

WO 02/23319 A1

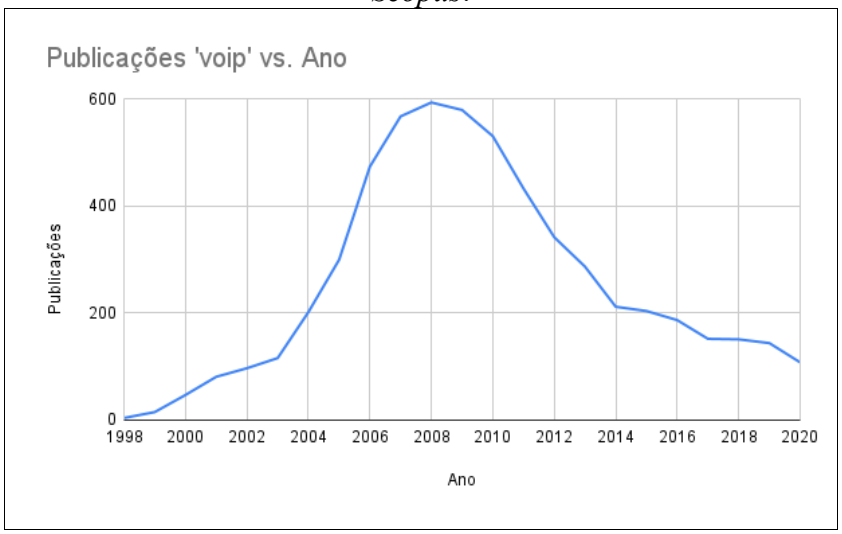
Fonte: Recuperado a partir do *Google Patents*®

Uma busca na base *Scopus*, restringindo às publicações em conferências, com a palavra '*voip*' nos campos '*title*', '*abstract*' e '*keywords*' retorna 5.898 documentos[35]. Ordenando de maneira crescente a partir da data mais antiga, tem-se uma publicação de 1998 como a primeira menção de *'voip'*. Observando a quantidade de publicações ao decorrer dos anos, como mostra o gráfico da Figura 31, percebe-se o crescimento significativo na quantidade de publicações a partir de 2003, mesmo ano que é tido como marco da disseminação da tecnologia pelo surgimento do Skype® e democratização do computador pessoal. Após 2008, o declínio no

[35] Busca realizada em 30 de novembro 2021.

número de publicações a partir da palavra '*voip*' ocorre, provavelmente, pela maturidade da tecnologia, tal que, pesquisas subsequentes se dedicam a termos mais específicos da área. Contudo, este não seria um fator preocupante para o modelo, pois antes do declínio, já teria sido possível sugerir a palavra como sinal fraco devido a sua prévia ascensão. Neste sentido, o modelo proposto nesta tese teria a capacidade de identificar a tecnologia VoIP como um sinal fraco a cerca dos anos 2000, alguns anos antes de se mostrar uma tendência promissora.

Figura 31 – Gráfico das publicações em conferências a partir da palavra '*voip*' na base *Scopus*.



Fonte: Elaborado pela autora

A relevância da obtenção dessa informação da tecnologia VoIP com antecedência, como um sinal fraco, consiste no valor de mercado para tecnologias disruptivas. Segundo Li, Porter, Suominen (2018), a descontinuidade tecnológica é representada em uma nova tecnologia que resulta na obsolescência de um produto ou serviço. Desse modo, o uso da *Internet* e do protocolo TCP/IP (*Transmission Control Protocol/ Internet Protocol*) para transportar chamadas de voz, representa uma descontinuidade tecnológica para as companhias telefônicas tradicionais e oferece uma oportunidade para empreender no ramo de telefonia com relativamente pouco investimento (SARITAS; SMITH, 2011). Logo, a tecnologia VoIP como sinal fraco se mostraria como uma informação notável para os processos de planejamento estratégico.

5.2 OPERACIONALIZAÇÃO DO MODELO

Nesta subseção está documentada uma simulação da implementação computacional do modelo proposto. Para o desenvolvimento deste cenário de estudo, utilizou-se o ambiente em nuvem Colab® da Google®. Este recurso, de plataforma como serviço (*Platform as a Service - PaaS*), disponibiliza simultaneamente o ambiente e a infraestrutura necessária para o desenvolvimento de *notebooks Python*®. Além disso, oferece também máquinas virtuais com arquiteturas CPU[36], GPU[37] e TPU[38] adaptadas para o desenvolvimento de modelos de *deep learning* como o modelo de linguagem BERT.

5.2.1 Etapa 1 - Elaborar o Conjunto de Dados

A estruturação do conjunto de textos (*dataset*) utilizado pelo modelo depende da recuperação dos documentos, extração da informação, pré-processamento dos dados e disponibilização do *corpus*. Tarefas descritas a seguir.

(1.1) Recuperação e Extração da Informação

Como fonte de conhecimento científico, bases de dados como o S2ORC®: *The Semantic Scholar Open Research Corpus* são importantes fontes para publicações científicas. A coleta de dados em bases de conhecimento disponíveis *online* podem ser realizadas por diferentes métodos e ferramentas, desde APIs oficiais até *softwares* externos como estratégias de *web scrapping* que coletam dados de plataformas. Idealmente, os processos dessa etapa seriam automatizados, coletando todas as publicações em conferências de cunho tecnológico de um determinado ano. Todavia, convém observar que analisar a efetividade computacional da forma como os dados são coletados, acessados ou armazenados está além do escopo desta pesquisa. Isto posto, para fins demonstrativos, sem comprometer a implementação do modelo, o conjunto de dados foi elaborado a partir da coleta manual das publicações em conferências indexadas apenas na base de publicações *Scopus* restritas à palavra de interesse *'voip',* nos campos '*title*', '*abstract*' e '*keywords*'.

Como resultado da busca, ordenando os documentos de maneira crescente a partir da data mais antiga, encontrou-se o seguinte volume de publicações para os primeiros 5 anos, como mostra a Tabela 2. Pontua-se que a eliminação de publicações duplicadas ou inconsistentes foi

[36] CPU: Central Processing Unit
[37] GPU: Graphical Processing Unit
[38] TPU: Tensor Processing Unit

realizada em simultâneo à coleta e organização dos arquivos de documentos, pois foi executado manualmente. Contudo, em um processo automatizado esta rotina de averiguação seria implementada sem complicações.

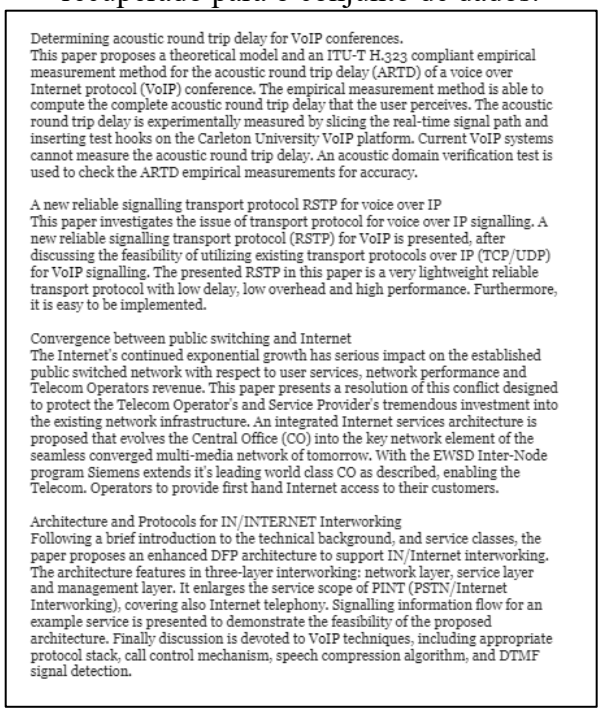
Tabela 2 – Relação número de publicações por ano de pesquisa.

Ano	Número de publicações
1998	4
1999	15
2000	46
2001	81
2002	96

Fonte: Elaborado pela autora

Foi considerado para a exemplificação esse intervalo de tempo (1998-2002) considerando o início da aparição da palavra '*voip*' nas publicações de congresso (1998) e na base de patentes (2001), bem como, representa a primeira inclinação de crescimento como visto no gráfico da Figura 31. O texto coletado de cada publicação foi organizado em arquivos .rtf, separadamente para cada ano. Na Figura 32, por exemplo, apresenta-se os 4 títulos e 4 resumos das 4 publicações do ano de 1998.

Figura 32 – Documento relativo ao ano 1998 como exemplo de arquivo .rtf contendo o texto recuperado para o conjunto de dados.



Determining acoustic round trip delay for VoIP conferences.
This paper proposes a theoretical model and an ITU-T H.323 compliant empirical measurement method for the acoustic round trip delay (ARTD) of a voice over Internet protocol (VoIP) conference. The empirical measurement method is able to compute the complete acoustic round trip delay that the user perceives. The acoustic round trip delay is experimentally measured by slicing the real-time signal path and inserting test hooks on the Carleton University VoIP platform. Current VoIP systems cannot measure the acoustic round trip delay. An acoustic domain verification test is used to check the ARTD empirical measurements for accuracy.

A new reliable signalling transport protocol RSTP for voice over IP
This paper investigates the issue of transport protocol for voice over IP signalling. A new reliable signalling transport protocol (RSTP) for VoIP is presented, after discussing the feasibility of utilizing existing transport protocols over IP (TCP/UDP) for VoIP signalling. The presented RSTP in this paper is a very lightweight reliable transport protocol with low delay, low overhead and high performance. Furthermore, it is easy to be implemented.

Convergence between public switching and Internet
The Internet's continued exponential growth has serious impact on the established public switched network with respect to user services, network performance and Telecom Operators revenue. This paper presents a resolution of this conflict designed to protect the Telecom Operator's and Service Provider's tremendous investment into the existing network infrastructure. An integrated Internet services architecture is proposed that evolves the Central Office (CO) into the key network element of the seamless converged multi-media network of tomorrow. With the EWSD Inter-Node program Siemens extends it's leading world class CO as described, enabling the Telecom. Operators to provide first hand Internet access to their customers.

Architecture and Protocols for IN/INTERNET Interworking
Following a brief introduction to the technical background, and service classes, the paper proposes an enhanced DFP architecture to support IN/Internet interworking. The architecture features in three-layer interworking: network layer, service layer and management layer. It enlarges the service scope of PINT (PSTN/Internet Interworking), covering also Internet telephony. Signalling information flow for an example service is presented to demonstrate the feasibility of the proposed architecture. Finally discussion is devoted to VoIP techniques, including appropriate protocol stack, call control mechanism, speech compression algorithm, and DTMF signal detection.

Fonte: Elaborado pela autora

(1.2) Pré-processamento

Para o pré-processamento do conteúdo textual dos documentos realizou-se a uniformização das palavras para caixa baixa e a remoção de caracteres especiais e numerais. Como também, os ajustes particulares para uso do modelo BERT, isto é, a inclusão dos *tokens* especiais e separação das sentenças. Bibliotecas de NLP tais como a NLTK® e spaCy, escritas em *Python*, são ferramentas úteis para essa etapa. Na Tabela 3, verifica-se como o *dataset* de cada ano ficou constituído.

Tabela 3 – Relação do número de sentenças por *dataset* e ocorrências da palavra *'voip'*.

Ano	Número de sentenças	Ocorrências da palavra *'voip'*
1998	24	7
1999	111	23
2000	279	79
2001	568	161
2002	704	195

Fonte: Elaborado pela autora

(1.3) Disponibilizar *corpus*

O *corpus* é composto por arquivos .rft, armazenados e organizados cronologicamente por ano, contendo como *dataset* o conjunto de títulos e resumos, extraídos e pré-processados, prontos para uso. Após isso, são armazenados no formato JSON (J*avaScript Object Notation*) em banco de dados orientados a documentos, permitindo que consultas em textos completos sejam realizadas.

5.2.2 Etapa 2 - Extrair e Representar Palavras do Conjunto de Dados

Neste momento, como o modelo BERT foi eleito para obtenção dos *embeddings*, as decisões descritas aqui dizem respeito às características e condições deste modelo. No entanto, tem-se ciência que futuramente existirão outras abordagens e outras decisões e ajustes serão necessários.

Dito isso, deve-se primeiramente realizar o *download* do modelo BERT *base uncased* acessível via *Python* (Figura 33). Na etapa anterior o texto foi preparado para o modelo BERT

separando as sentenças[39]. Agora, para obter a representação vetorial das palavras, tem-se a etapa de *token*ização e em seguida a extração de atributos resultando na obtenção dos *embeddings*.

Figura 33 – Importação da biblioteca do modelo BERT, spaCy e outras de suporte.

```
[ ]  !pip install transformers striprtf torch pandas tqdm
     !pip install -U spacy[cuda110]

[ ]  import torch, gc
     import pandas as pd
     import numpy as np
     import spacy
     from sklearn.feature_extraction.text import TfidfVectorizer
     from striprtf.striprtf import rtf_to_text
     import re
     import random
     import tensorflow as tf
     from transformers import BertTokenizer, BertForPreTraining, BertModel
     from tqdm import tqdm
```

Fonte: Elaborado pela autora

O modelo BERT executa internamente a *token*ização com a implementação do algoritmo *WordPieces* (WU *et al.*, 2016). Relembrando, ocorre o reconhecimento do texto como palavras ou subpalavras de acordo com seu vocabulário de *tokens* pré-treinados, atribuindo-se um *token-id* único para cada unidade. Esse processo é exemplificado na Figura 34, na qual pode-se ver que ao receber a sentença, o *tokenizer* retorna os *tokens* presentes na sentença e seus respectivos identificadores (*ids*). Neste exemplo, a palavra *'interworking'* não é reconhecida pelo dicionário e, por esse motivo, é fragmentada em dois *tokens* *'inter'* e *'##working'* com *ids* [6970] e [21398], respectivamente.

Uma vez que os ajustes no âmbito de texto são concluídos, a próxima etapa para obtenção dos *embeddings* é a realização das tarefas padrão do modelo BERT de linguagem mascarada (ML) e previsão da próxima sentença (NSP) no conjunto de texto específico pré-estabelecido, ou seja, a utilização do BERT para extração de atributos a partir do *corpus* elaborado.

Para tanto, o *dataset* é dividido em um número de épocas, definido conforme a quantidade de iterações que serão realizadas durante a execução da tarefa de extração de atributos do BERT. Para este cenário de estudo, o número de épocas foi definido empiricamente, verificando quantas iterações eram necessárias para ajustar o *embedding* da palavra nova, de modo que seu significado se aproximasse de palavras da mesma área.

[39] Frase de exemplo é uma sentença extraído do *dataset* de 1998.

Verificou-se assim que 10 épocas apresentavam um resultado satisfatório. Findada esta etapa de configurações, o modelo está pronto para a obtenção dos *embeddings* a partir do *corpus* pré-estabelecido.

Figura 34 – Exemplo de entrada e saída do processo de *token*ização *WordPiece* do modelo BERT.

```
!pip install transformers striprtf torch pandas tqdm
from transformers import BertTokenizer

# Load the tokenizer
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

# Define a sample sentence
sentence = "the architecture features in three layer interworking"

# Tokenize the sentence
token_ids = tokenizer.encode(sentence, add_special_tokens=True)

# Print the token ids and their corresponding tokens
for i, token_id in enumerate(token_ids):
    print("Token:", tokenizer.convert_ids_to_tokens([token_id])[0], "\tId:", token_id)
```

```
Token: [CLS]    Id: 101
Token: the      Id: 1996
Token: architecture     Id: 4294
Token: features         Id: 2838
Token: in       Id: 1999
Token: three    Id: 2093
Token: layer    Id: 6741
Token: inter    Id: 6970
Token: ##working        Id: 21398
Token: [SEP]    Id: 102
```

Fonte: Elaborado pela autora

A tarefa ocorre tal que cada um dos *tokens* do vocabulário possui um vetor de representação atribuído na etapa de *token*ização de acordo com o pré-treinamento do modelo. As 768 posições deste *embedding* são modificadas durante a execução das tarefas padrão conforme a contextualização de cada palavra no *dataset* pré-estabelecido. Assim, ao final da etapa de extração de atributo, os *tokens* possuem um novo *embedding* ajustado em relação ao *dataset*. Por fim, cabe ressaltar que a tarefa de extração de atributos será realizada na etapa seguinte após a identificação e inserção da palavra nova no vocabulário no modelo BERT. Tal ajuste na sequência efetiva das etapas do modelo proposto ocorre pela especificidade da abordagem para *embeddings* escolhida, ao passo que o modelo proposto é elaborado de forma mais abrangente de diretrizes.

Como saída desta etapa tem-se o texto *token*izado e com seus respectivos *tokens-id* associados, porém, ainda sem ajuste do vetor dos *embeddings* para a representação semântica contextual do texto específico.

5.2.3 Etapa 3 - Identificar Sinais Novos

A identificação de sinais novos depende da descoberta das novas palavras e obtenção de seus *embeddings*. Neste caso, executa-se apenas a situação em que novas palavras remetem à palavras inéditas. Para realizar essa identificação recorreu-se ao *token*izador provido pela biblioteca spaCy e a uma rotina inspirada no algoritmo exBERT (TAI *et al*. 2020). Em resumo, a descoberta e obtenção do *embeddings* de novas palavras ocorre por meio das seguintes etapas:

1. Descobrir lista de palavras novas - interseção das palavras marcadas como desconhecidas pelo spaCy e palavras particionadas pelo modelo BERT;
2. Adicionar palavras novas ao vocabulário do *token*izador do BERT;
3. Obter seus *embeddigns* - implementação do modelo BERT na função de Extração de Atributos, isto é, execução do modelo em suas tarefas padrão com o *corpus* de documentos pré-estabelecido referentes às palavras novas.

(3.1) Descobrir lista de palavras novas

A descoberta da palavra ocorre pela intersecção das palavras marcadas como desconhecidas pelo spaCy e palavras particionadas pelo modelo BERT. As Figuras 35, 36 e 37 exemplificam a descoberta da palavra '*voip*' a partir da sentença *"determining acoustic round trip delay for voip conferences"* que representa o título da primeira publicação de 1998. Convém observar, na Figura 35 que a palavra '*voip*' foi alterada para '*voixp*' para poder demonstrar no *token*izador spaCy sua capacidade de identificar uma palavra que está fora de seu vocabulário, tal alteração se fez necessária uma vez que a palavra '*voip*', atualmente, faz parte de seu vocabulário.

Para a identificação das palavras pelo *token*izador spaCy recorre-se ao atributo "*.is_oov*" que indica se o *token* está fora do vocabulário (OOV) usado pelo spaCy. Caso o *token* for OOV, o atributo "*.is_oov*" retornará '*True*'; caso contrário, retornará '*False*', certificando que a palavra consta em seu vocabulário. Dessa forma, ao identificar as palavras OOV pode-se realizar o tratamento adequado dessas palavras, como substituí-las por um *token* "desconhecido" ('*UNK*'). Observa-se no exemplo como saída a consulta '*True/False*' se a palavra existe no vocabulário, mostrando apenas a palavra *'voixp'* como *'True'*, ou seja, não existe no vocabulário do spaCy. Sendo assim, na saída da sentença *token*izada, esta palavra desconhecida é marcada como '*UNK*'.

Figura 35 – *Token*ização da sentença pelo spaCy marcando palavras fora do vocabulário como '*UNK*'.

```
!python -m spacy download en_core_web_md
import spacy

# Tokenization using spaCy
nlp = spacy.load("en_core_web_md")
tokens = nlp("determining acoustic round trip delay for voixp conferences")

# token.is_oov is used to identify OOV words.
for token in tokens:
    print(token.text, token.is_oov)
# Create a tokenizer with 'UNK' as the out-of-vocabulary token.
unk_token=[]
for token in tokens:
    if not token.is_oov:
        unk_token.append(token.text)
    else:
        unk_token.append('UNK')
print(unk_token)
```

```
determining False
acoustic False
round False
trip False
delay False
for False
voixp True
conferences False
['determining', 'acoustic', 'round', 'trip', 'delay', 'for', 'UNK', 'conferences']
```

Fonte: Elaborado pela autora

A Figura 36, traz um comparativo da saída para a mesma sentença nos dois *tokenizers*. Na Figura 36 (a), verifica-se a saída do modelo BERT separando a palavra '*voip*' como ['*vo*', '*##ip*'] e na Figura 36(b) tem-se o mesmo *token* classificado como desconhecido pelo spaCy. Dessa maneira, pode-se realizar uma comparação capaz de associar as palavras marcadas como 'UNK' pelo spaCy com as palavras particionadas pelo *tokenizer* do modelo BERT e, por conseguinte, identificar palavras novas.

Figura 36 – Comparativo de sentença com palavra desconhecida *token*izada pelo BERT (a) e pelo spaCy (b).

(a)

```
['determining', 'acoustic', 'round', 'trip', 'delay', 'for', 'vo', '##ip', 'conferences']
```

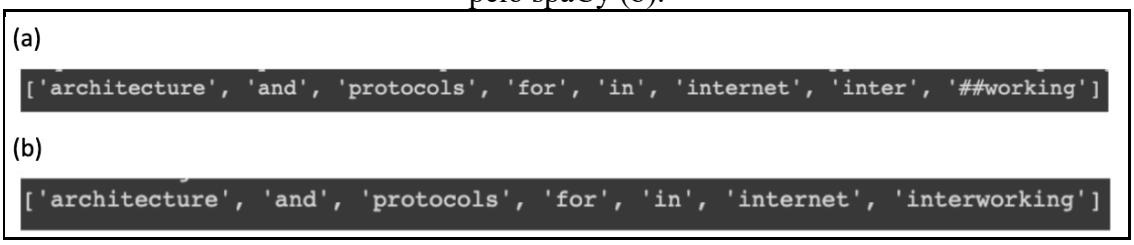
(b)

```
['determining', 'acoustic', 'round', 'trip', 'delay', 'for', 'UNK', 'conferences']
```

Fonte: Elaborado pela autora

Reforçando, a utilidade do *token*izer spaCy reside no seu uso como um filtro, pois o modelo BERT não particiona apenas palavras desconhecidas (OOV), mas particiona todas aquelas palavras que beneficiem seu processamento, como radicais, prefixos, sufixos e outros elementos textuais. Como exemplo, pode-se observar na Figura 37, em que a sentença *"the architecture features in three layer interworking"* não possui palavras desconhecidas pelo vocabulário spaCy. No entanto, o modelo BERT particiona a palavra *'interworking'* em dois *tokens* [*'inter', '##working'*]. Nesse sentido, por mais que o modelo BERT particione a palavra, isto não implica que seja uma palavra nova de interesse para o modelo proposto. Por essa razão, o uso da interseção com a *token*ização do spaCy se mostra adequada para identificar novas palavras.

Figura 37 – Comparativo de sentença sem palavra desconhecida *token*izada pelo BERT (a) e pelo spaCy (b).



Fonte: Elaborado pela autora

Na Figura 38 pode-se verificar uma rotina implementada de modo a ilustrar tal processo. A função, ao encontrar um *token* '*UNK*' na *token*ização oferecida pelo spaCy, busca na *token*ização provida pelo BERT a palavra particionada a qual corresponde. A função então identifica a 'palavra desconhecida' como o *token* anterior ao *token* marcado por "##" (que indica que a palavra foi particionada), e concatena esses *tokens* até que o próximo *token* na leitura da lista não inicie mais pela dupla cerquilha, indicando que finalizou a quantidade de *tokens* relativos às subpalavras que foram particionadas. Ou seja, 'UNK' = 'vo' + '##ip' = 'voip'. Evidentemente, uma aplicação para um grande volume de dados necessita de ajustes no código e decisões mais robustas de programação. Uma vez identificadas, essas palavras desconhecidas são armazenadas em uma lista como "Sinais Novos".

Figura 38 – Exemplifica como a palavra desconhecida ‘UNK’ seria encontrada dada as sentenças *token*izadas pelo BERT e pelo spaCy.

```
#match the UNK token with the new word in BERT

def match_unk(list_BERT, list_SPACY):
    for i, word in enumerate(list_SPACY):
        if word == 'UNK':
            matching_word = ''
            for j in range(i+1, len(list_BERT)):
                if list_BERT[j].startswith('##'):
                    matching_word = list_BERT[i]+ list_BERT[j][2:]
                else:
                    break
            return f"the UNK word is: '{matching_word}'"
    return "No match found"


list_BERT = ['determining', 'acoustic', 'round', 'trip', 'delay', 'for', 'vo', '##ip', 'conferences']
list_SPACY = ['determining', 'acoustic', 'round', 'trip', 'delay', 'for', 'UNK', 'conferences']

print(match_unk(list_BERT, list_SPACY)) # Output: the UNK word is: 'voip'
```

```
the UNK word is: 'voip'
```

Fonte: Elaborado pela autora

Uma observação a ser realizada nessa etapa diz respeito à desconsideração de palavras com menos de quatro letras a fim de eliminar possíveis ruídos, uma vez que poucas palavras significativas apresentam três caracteres ou menos.

(3.2) Incluir palavra nova no vocabulário BERT

Uma vez identificadas as palavras novas, o próximo passo é obter seus *embeddings*. Para tanto, é necessário adicionar as palavras novas ao vocabulário do *token*izador do BERT, para que a palavra seja compreendida por um único *token* pelo modelo e possibilitar sua análise temporal. Neste sentido, uma rotina inspirada na expansão exBERT (TAI *et al.*, 2020), foi implementada adicionando a nova palavra como único *token* ao vocabulário do BERT.

O modelo BERT oferece duas maneiras de se realizar inclusões de novas palavras ao seu vocabulário. A primeira permite a inserção direta da OOV no arquivo de vocabulário presente no diretório raiz do BERT. Este documento, o ‘*vocab.txt*’, possui todos os *tokens* de palavras conhecidas pelo modelo, incluindo subpalavras, numerais, símbolos e caracteres especiais. O arquivo possui ainda aproximadamente mil espaços não utilizados, nomeados como ‘*[unusedn]*’, reservados para a inclusão de novas palavras, bastando substituir essas lacunas pelas palavras a serem adicionadas. Por padrão, o modelo ignora os *tokens* do vocabulário que estejam escritos entre colchetes, por esse motivo, os espaços não ocupados do vocabulário não afetam o desempenho do modelo nas tarefas de NLP.

Outra forma de inclusão de OOV ao modelo BERT é por meio de uma função nativa do *tokenizer*, o método '*add_tokens*', que recebe como parâmetro os termos a serem adicionados ao vocabulário. Este método, todavia, não realiza alterações no arquivo do vocabulário original, '*vocab.txt*', mas cria um arquivo anexo de configuração do tipo JSON, que armazena as novas palavras e seus respectivos *ids*. O método '*add_tokens'*, apesar de ser nativo, faz uma alteração no número de linhas do vocabulário, exigindo que seja rebalanceado o tamanho da matriz *token-embeddings* do modelo, sendo portanto, uma estratégia recomendada nos casos em que é necessária a inclusão de um grande número de palavras.

Diante dessas possibilidades, visando a exemplificação do modelo, optou-se pela inclusão direta da palavra ao arquivo de vocabulário (Figura 39). Importante destacar que esta primeira etapa permite apenas que o BERT reconheça a nova palavra como um *token*, não sendo atribuído nenhum tipo de *embedding* significativo ou contextual . A palavra *'voip'* foi incluída no vocabulário do modelo BERT utilizando o espaço não utilizado *'[unused992]'*, presente no arquivo *vocab.txt.*

Figura 39 – Adição da palavra nova *'voip'* no vocabulário do modelo BERT.

```
[ ]  new_tokens = []
     new_tokens.append('voip')
     new_tokens

     ['voip']

[ ]  #carga do vocabulário original do BERT
     sentences = []
     with open(output_dir + '/vocab.txt', 'r') as arq:
       sentences = str.split(arq.read(), sep='\n')

     #adição das palavras novas à lista de vocabulário
     for i in range(0,len(new_tokens),1):
       sentences[998-i] = new_tokens[i]
```

Fonte: Elaborado pela autora

Com o modelo reconhecendo a palavra como um único *token*, prossegue-se para a obtenção dos *embeddings*.

(3.3) Obter *embeddings* - BERT *Feature Extraction*

A fim de obter os *embeddings* das palavras, utiliza-se o modelo BERT em sua capacidade de extração de atributos *(Feature Extration*). Isso significa a utilização do modelo

BERT em suas tarefas padrão (ML e NSP), com o *corpus* de documentos específico, pré-definido, referentes às novas palavras para poder obter seus *embeddings*. Por essa capacidade de ajustar os *embeddings* das palavras com um *corpus* dedicado, os *embeddings* fornecidos pelo modelo BERT são denominados dinâmicos.

Pontualmente, para a palavra nova e para cada ano, o *embedding* é calculado e acumulado, isto é, a interação do ano seguinte parte do *embedding* obtido no ano anterior. Essa decisão diz respeito ao aprimoramento do *embedding* da palavra nova que inicia com seu vetor sem representação contextual. Neste exemplo, a primeira iteração parte da palavra *'voip*' inserida no vocabulário BERT e com o *dataset* de 1998. Este processo de extração de atributos gerou *embeddings* ajustados conforme o contexto e as ocorrências das palavras nos artigos de 1998. Os *embeddings* ajustados foram então usados como partida para o *dataset* de 1999, incorporando assim o aprendizado das ocorrências das palavras deste ano. Esse processo foi repetido sequencialmente, com os *embeddings* de 1999 sendo ajustado para 2000, 2000 para 2001 e 2001 para 2002. Assim, a cada novo *dataset,* o cálculo dos *embedding* para a palavra nova, no exemplo sendo a palavra *'voip*', acumulou seu aprendizado com os anos anteriores.

Esse é um aspecto que requer atenção, pois é esperado que as primeiras tentativas para obter o *embedding* da palavra nova não sejam significativas, uma vez que a quantidade de ocorrência da palavra é baixa (como mostra a Tabela 3). Por esse motivo, se torna importante acumular os *embeddings* progredindo a capacidade do modelo de representar vetorialmente o contexto da palavra nova. À medida que o número de publicações aumenta com o passar do tempo, aumenta o número de sentenças e, consequentemente, a quantidade de ocorrências da palavra de interesse. Como o *embedding* do *token* é ajustado no modelo sempre que há uma ocorrência da palavra correspondente no *dataset*, quanto maior for o número de ocorrências da palavra, mais consolidada é sua representação vetorial semântica contextual.

Como saída desta etapa tem-se uma lista das palavras novas identificadas, consideradas 'sinais novos', e seus *embeddings* respectivamente calculados.

5.2.4 Etapa 4 - Elaborar Rede de Palavras

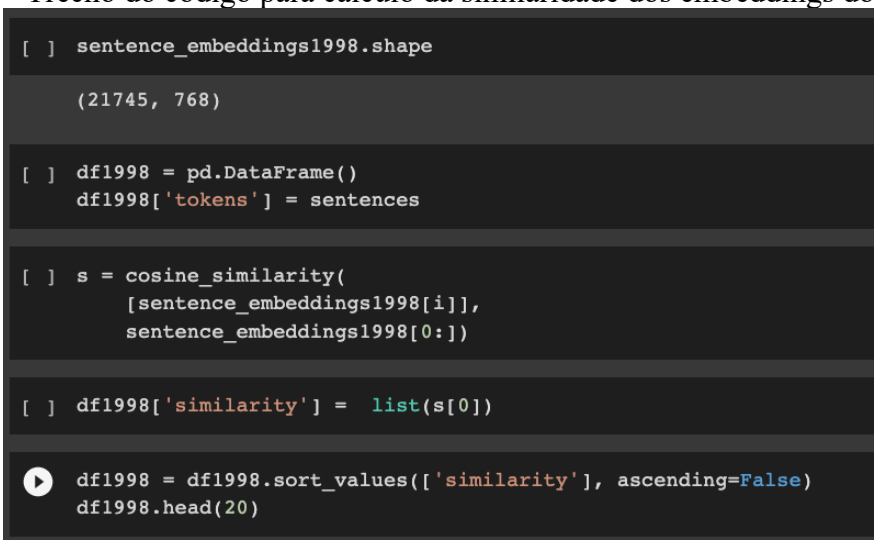
Nesta etapa objetiva-se em conceber as redes de palavras como meio para monitorar temporalmente a evolução do *embedding* da palavra nova. Os *embeddings* enquanto representações vetoriais densas das palavras são espacialmente posicionados, no entanto, não é

explícita a relação e conexão entre as palavras. Assim, tem-se a tarefa de explicitar a relação entre as palavras e exibir a rede.

(4.1) Calcular a similaridade

A relação entre os nós da rede é estabelecida pelo cálculo de similaridade de cossenos entre as palavras para cada ano (Figura 40).

Figura 40 – Trecho do código para cálculo da similaridade dos embeddings do ano 1998.



Fonte: Elaborado pela autora

A Figura 41 apresenta a similaridade da palavra *'voip'* com as 20 palavras mais próximas ao longo de diferentes anos. Na primeira posição encontra-se a própria palavra de estudo, pois sendo uma palavra igual a ela mesma a similaridade é máxima (1). '*None*' é o ponto de partida onde a palavra '*voip*' foi apenas adicionada ao vocabulário, sendo o ano de 1998 o primeiro ajuste contextual para obtenção dos *embeddings*.

Nos dois primeiros anos de ajuste, 1998 e 1999, pode-se notar valores altos de similaridade (acima de 0.90), no entanto, as palavras próximas não descrevem o contexto da palavra '*voip'*. Este é um resultado esperado, uma vez que existe um baixo número de ocorrências da palavra '*voip'* nos *datasets* desses anos (Tabela 3), não sendo suficientes para produzir uma representação aderente ao contexto semântico de maneira satisfatória.

Na interação do ano 2000 inicia-se a convergência contextual, em que as similaridades mais altas começaram a apontar para palavras presentes no mesmo campo semântico de '*voip'*. Contudo, nota-se que o valor das similaridades reduziu. Tal característica reflete que a representação se torna menos genérica e se aproxima de palavras similares contextualmente.

Ao passo que, embora individualmente as similaridades reduzam, existe uma relação entre todas as palavras listadas, elas descrevem um contexto. Ou seja, essa lista de palavras representa um subconjunto de nós mais interconectados em comparação à lista de palavras dos anos anteriores.

Para os dois últimos anos de ajuste, nota-se que os resultados começaram a apresentar uma tendência de estabilização. Isto é, pode-se inferir que o cálculo do *embedding* da palavra nova *'voip'* é capaz de realizar sua representação semântica contextual.

Esse processo de observação da capacidade de representação do *embedding* da palavra nova é um passo importante para o sucesso do modelo. Entende-se a aproximação do significado da palavra nova com outras palavras que descrevem o mesmo contexto, como um critério determinante para se ter um *embedding* significativo. Afinal, para uma boa representação da palavra, seu *embedding* deve apresentar uma alta similaridade de cosseno com *embeddings* de palavras próximas por apresentarem características em comum.

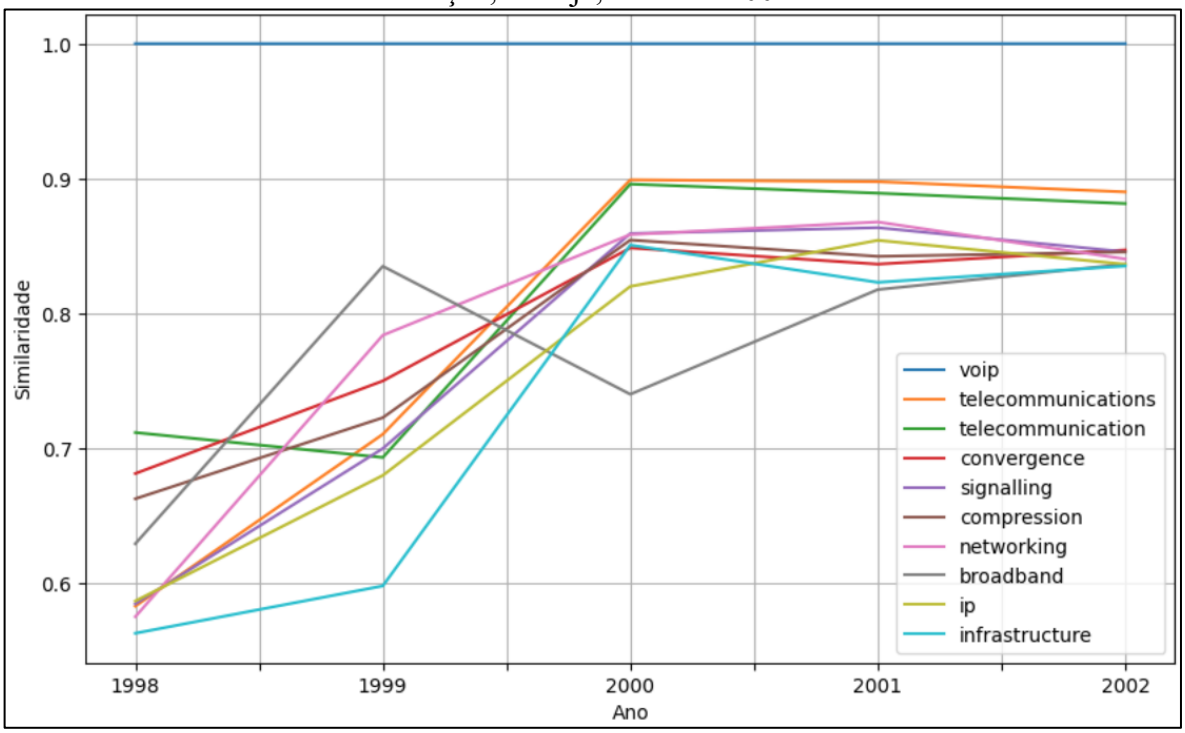
Figura 41 – Similaridade da palavra *'voip'* com as 20 palavras mais semelhantes de cada ano.

None		1998		1999		2000		2001		2002	
tokens	similarity	tokens	similarity	tokens	similarity	tokens	similarity	tokens	similarity	tokens	similarity
voip	1.000000	voip	1.000000	voip	1.000000	voip	1.000000	voip	1.000000	voip	1.000000
appellate	0.961228	clergyman	0.925057	eurasian	0.964764	telecommunications	0.899122	telecommunications	0.897772	telecommunications	0.890165
rq	0.953710	commuted	0.924749	constituent	0.961794	telecommunication	0.895975	telecommunication	0.889255	telecommunication	0.881561
blackish	0.952081	harassed	0.918772	wavelengths	0.961767	integration	0.866948	networking	0.867918	convergence	0.847171
populated	0.951243	uniformly	0.918144	palatine	0.961276	signalling	0.859364	switching	0.865531	signalling	0.845802
buren	0.950665	impoverished	0.916904	sugarcane	0.960455	networking	0.858580	signalling	0.863687	compression	0.845660
endeavors	0.949535	incomes	0.916211	unopposed	0.959830	terminals	0.856625	ip	0.854326	networking	0.840471
gabled	0.945545	overgrown	0.915642	rye	0.959088	switching	0.856350	compression	0.842433	broadband	0.836684
stemming	0.943198	gabled	0.915257	discourage	0.957833	multimedia	0.855219	multimedia	0.842307	ip	0.836594
independents	0.942809	stifled	0.913539	forerunner	0.956381	networks	0.854971	convergence	0.836696	infrastructure	0.835289
incomes	0.942628	endeavors	0.912796	recreated	0.956073	compression	0.854536	networks	0.833731	switching	0.835028
overgrown	0.942194	methodist	0.912241	openings	0.955791	computing	0.853231	wireless	0.824097	multimedia	0.829543
propped	0.936587	relies	0.912169	irregular	0.954317	standardization	0.851669	infrastructure	0.823177	deployment	0.828144
overs	0.935934	celtics	0.911279	reinstated	0.954296	infrastructure	0.850811	integration	0.818481	feasibility	0.826214
staten	0.934167	biceps	0.910162	relaunched	0.954039	protocols	0.850782	protocols	0.818335	integration	0.825921
geese	0.934161	confines	0.909410	sighted	0.953812	routing	0.849096	broadband	0.817818	upstream	0.823656
supplemented	0.933936	restrictive	0.909196	triangular	0.953804	optimization	0.848806	ubiquitous	0.815054	optimization	0.823055
stifled	0.933589	undergoing	0.908943	inconsistent	0.953735	convergence	0.848691	sip	0.813307	internet	0.820820
relies	0.933513	vidhan	0.908860	bodied	0.953628	technologies	0.847643	optimization	0.810537	guarantees	0.815576
sentai	0.933225	tilted	0.908414	subtropical	0.953113	feasibility	0.845937	telecom	0.809548	routing	0.815076

Fonte: Elaborado pela autora

O gráfico expresso na Figura 42 demonstra este processo de progressão, para as 10 palavras consideradas mais similares ao longo dos 5 ajustes finos a partir da palavra de interesse *'voip*'. É notável a tendência de subida das similaridades ao longo dos dois primeiros ajustes, seguida de uma estabilização nos dois ajustes posteriores.

Figura 42 – Gráfico do avanço temporal das palavras mais próximas exibidas na última iteração, ou seja, o ano de 2002.



Fonte: Elaborado pela autora

Cabe mencionar que a escolha de '*voip*' como caso de estudo para nova palavra se mostrou ainda interessante por se tratar de um acrônimo. Como visto na seção 2.4, na formação de nomenclaturas, *'voip*' se diferencia do caso de '*graphene*', por exemplo, por ser uma palavra que deriva de uma palavra já existente (*graphite*) e sufixo conhecido da química, '*-ene'*, tal que seria esperado que o modelo fosse mais hábil para obter seu *embedding*. Nesse sentido, investigar o comportamento do *embedding* no caso do acrônimo '*voip*' e observar a capacidade de aproximar a palavra nova para o contexto ao qual ela se refere, se mostrou relevante. Esse processo de *embedding* para novas palavras com o modelo BERT encontra-se publicado (MACIEL; ARTESE; GONÇALVES, 2022) na conferência CONTECSI (*International Conference on Information Systems and Technology Management*).

(4.2) Exibir rede de palavras

Na subtarefa anterior foram calculadas as similaridades como critério de relação entre as palavras. Neste momento, estabeleceu-se um critério de corte (*threshold*), tal que, apenas palavras que apresentem similaridade acima deste limiar tem suas conexões representadas na rede, onde cada palavra, isto é, seu *embedding* é um nó da rede.

A Figura 43 apresenta uma função cujo propósito é calcular um *threshold* que pode ser usado para filtrar as similaridades mais fracas, menos significativas. O raciocínio envolvido é adicionar a similaridade mais alta a uma medida de tendência central (a mediana dos logaritmos) para obter um limite que considere a distribuição dos valores de similaridade do conjunto de dados. Tomando a mediana dos logaritmos, efetivamente está se calculando um valor de similaridade "típico" que pode ser adicionado ao valor máximo de similaridade para obter um limite razoável.

Figura 43 – Limite heurístico das similaridades de cosseno para um conjunto de dados.

```
[ ]  def get_threshold(df):
       log_list = []
       max_similarity = 0.0
       ctr = 0
       for index, row in df.iterrows():
         if (ctr > 0):
           if (ctr == 1):
             max_similarity = row.similarity
           log_list.append(np.log10(row.similarity))
         ctr = ctr + 1
       return max_similarity+np.median(log_list)
```

Fonte: Elaborado pela autora

Importante mencionar que a construção do grafo foi limitada devido a restrições computacionais. Considerando a ordem de tamanho dos grafos, números de nós e arestas, seria necessário um poder computacional especializado, portanto, foi necessária uma limitação instrumental para gerar essa operacionalização. Todavia, cuidados foram tomados para que essa demonstração do grafo simulasse o comportamento do grafo real, por meio da implementação de critérios de crescimento orgânico para o recorte apresentado, visando uma escala reduzida que fosse computacionalmente viável.

A Figura 44 apresenta a função *'generate_subgraph'* que constrói o grafo a partir de uma palavra específica, encontrando as palavras mais semelhantes conforme a similaridade dos cossenos, até um número máximo de níveis especificado por '*max_level*'. Toda vez que o valor da similaridade for maior ou igual ao limite, uma aresta é adicionada ao grafo e a função é chamada recursivamente para adicionar mais nós e arestas ao grafo até que o nível máximo seja atingido. O grafo é representado como uma lista de adjacências, onde cada nó é um *token* e as arestas representam a similaridade entre os nós. Assim, para cada ano produz-se um grafo representando as palavras semanticamente mais semelhantes de um *dataset*.

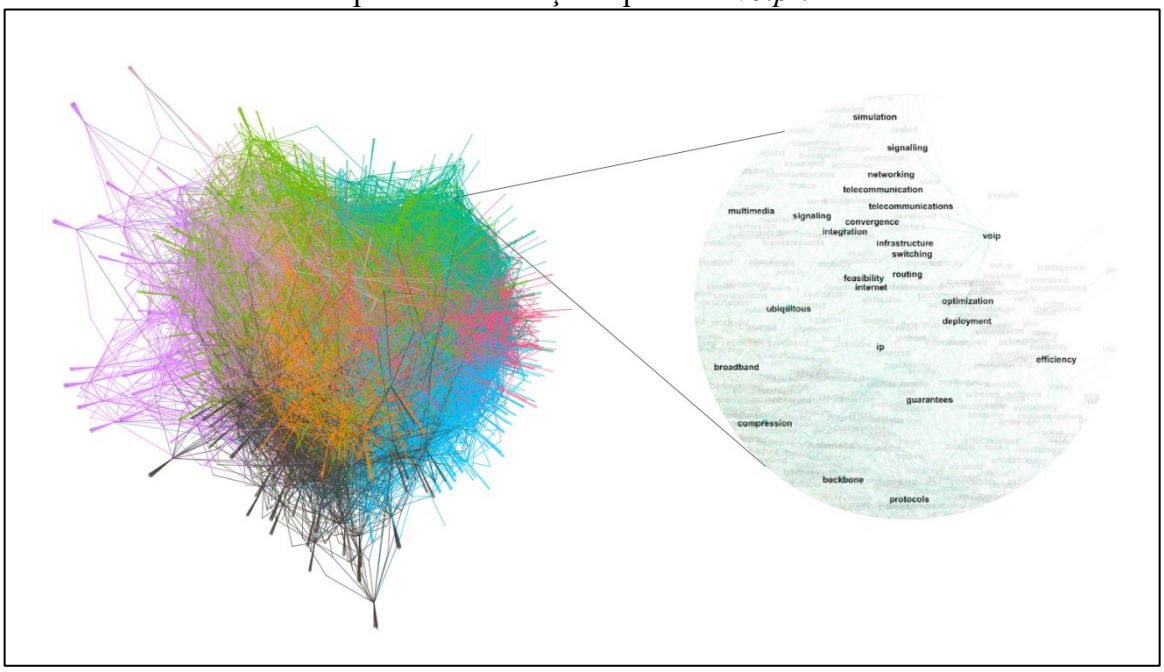
Figura 44 – Construção do grafo do *corpus* contendo a palavra *'voip'*.

```
def generate_subgraph_(token:str, specific_sentence, max_level:int, current_level:int, number_of_nodes:int, adjacency_list):

  if (current_level > max_level):
    return

  ctr = 0
  source = ''

  i = sentences.index(token)

  df = pd.DataFrame()
  df['token'] = sentences

  s = cosine_similarity(
      [specific_sentence[i]],
      specific_sentence[0:])

  df['similarity'] =  list(s[0])

  df = df.sort_values(['similarity'], ascending=False).head(new_number_of_nodes)

  threshold = get_threshold(df)

  for index, row in df.iterrows():
    if (ctr > 0):
      if (row.similarity >= threshold):
        store_relation(adjacency_list, source, row.token, row.similarity, 1)
        subgraph_new(row.token, specific_sentence, max_level, current_level + 1, number_of_nodes, adjacency_list)
      else:
        break;
    else:
      source = row.token
    ctr = ctr + 1
```

Fonte: Elaborado pela autora

Com o uso de *softwares* e bibliotecas de visualização pode-se exibir as redes elaboradas a cada recorte de tempo (*snapshots*). Bibliotecas visuais disponíveis em *Python* como *NetworkX*, podem ser uma alternativa para essa visualização. Na Figura 45, visualiza-se o grafo estático referente ao ano de 2002. A fim de transmitir uma noção de sua dimensão, mesmo esse grafo gerado como um recorte mínimo viável de processamento possui mais de 5 mil nós e aproximadamente 34 mil conexões. Em destaque pode-se notar as palavras que representam o contexto da palavra *'voip'*.

Figura 45 – Visualização do grafo gerado a partir do *dataset* do ano de 2002, com destaque para a vizinhança da palavra *'voip'*.



Fonte: Elaborado pela autora

5.2.5 Etapa 5 - Identificar Sinais Latentes

Nesta etapa o objetivo é monitorar e identificar uma taxa acelerada de crescimento (*burst*) nas relações no subgrafo que contém a palavra nova. Com isso em mente, algoritmos para análise temporal dinâmica de grafos como o "*Top-k Density Bursting Subgraph Finding*" (TDF) de Chu *et al*. (2019) atendem a necessidade. No entanto, a implementação deste tipo de análise, ou seja, encontrar subgrafos em redes temporais que variam em quantidade e tamanho, é um problema de difícil solução, inclusive, sendo um tópico em atual desenvolvimento e requer um alto poder computacional. À vista disso, utilizou-se o princípio do algoritmo temporal de modo estático, em outras palavras, foram consideradas métricas similares para inferir sobre o crescimento do subgrafo ano a ano.

Idealmente, nesta etapa os subgrafos de comportamento *burst* seriam identificados segundo um algoritmo e seria verificada a presença das palavras rotuladas como 'sinais novos' nestes subgrafos. No entanto, para ilustrar esta etapa do modelo optou-se por simular um subgrafo partindo da palavra de interesse '*voip*' e estimar seu crescimento temporalmente.

A função '*extract_subgraph*' (Figura 46) destina-se a extrair um subgrafo do grafo original relacionado à palavra de interesse '*voip'*. Utiliza-se a '*adjacency_list*' que contém informações sobre as relações entre os nós do grafo para obter o subgrafo. A função então

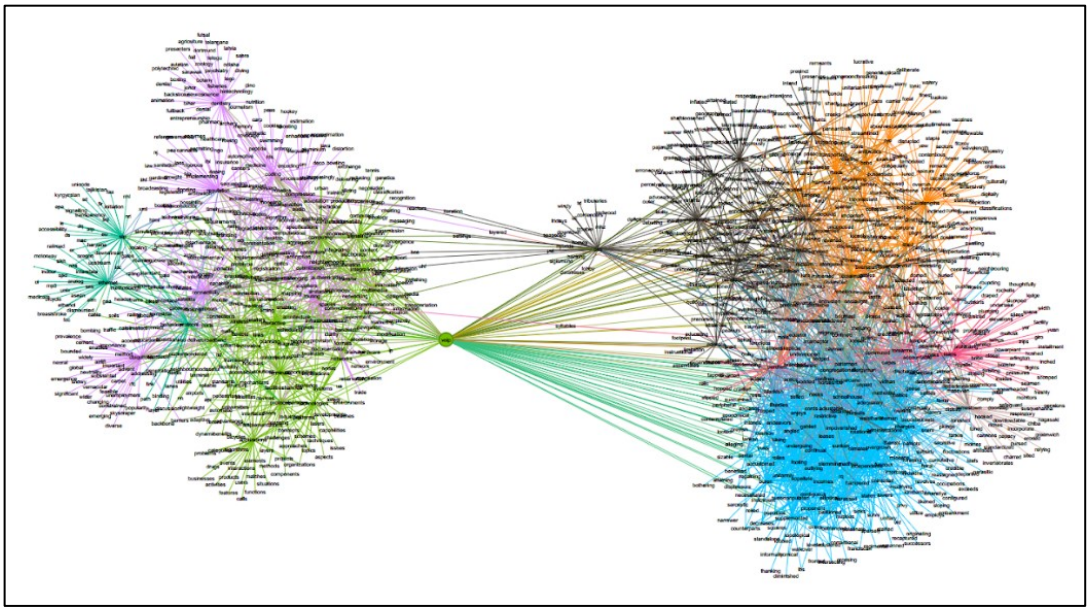
percorre a lista de adjacência e obtêm os nós diretamente conectados a '*voip*' (isso é referido como o 1º nível). Então, para cada um desses nós, a função obtêm os nós diretamente conectados a eles (isso é chamado de 2º nível). A saída é uma nova lista de adjacência (*'new_adjacency_list'*) que representa o subgrafo, contendo a relação entre os nós obtidos no 1º e 2º níveis. A Figura 47 apresenta o subgrafo extraído a partir do grafo acumulado de 2002.

Figura 46 – Extração do subgrafo em dois níveis a partir da palavra de interesse '*voip*'.

```
def extract_subgraph(adjacency_list:dict, term:str, type=2):
  new_adjacency_list = {}
  sl_term_list = []
  #1st level
  for source in adjacency_list:
    if (source == term):
      inner_hash = adjacency_list[source]
      for target in inner_hash:
        sl_term_list.append(target)
        store_relation(new_adjacency_list, source, target, inner_hash[target], 1)

  #2nd level
  if (type == 2):
    for sl_term in sl_term_list:
      for source in adjacency_list:
        if (source == sl_term):
          inner_hash = adjacency_list[source]
          for target in inner_hash:
            store_relation(new_adjacency_list, source, target, inner_hash[target], 1)

  return new_adjacency_list
```

Fonte: Elaborado pela autora

Figura 47 – Subgrafo obtido a partir da palavra '*voip*' extraído com base no grafo acumulado em 2002.



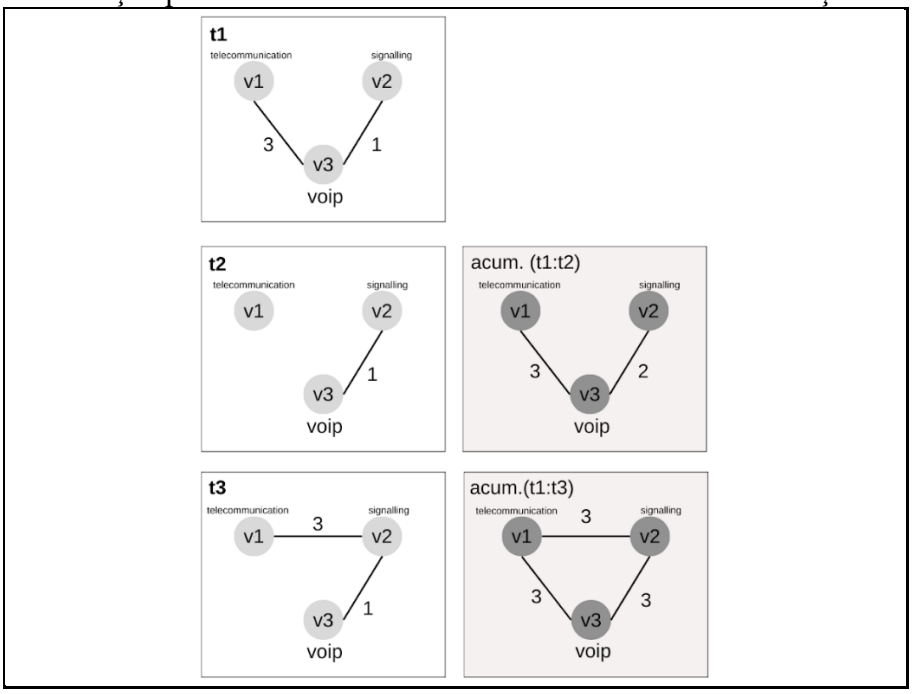
Fonte: Elaborado pela autora

O algoritmo TDF detecta o subgrafo de crescimento *burst* identificando o subconjunto de nós que maximiza a métrica de *Cohesiveness* do grafo no menor intervalo de tempo (seção 2.3.2.1). Essa é uma medida para inferir sobre quão fortemente são conectados entre si os nós de um subgrafo. Para essa métrica, o algoritmo tem como base a matriz de afinidade da rede, isso significa que considera a força das conexões (o peso das arestas) no cálculo (Equação 1). Dessa forma, para o algoritmo a intensificação da relação, isto é, o acúmulo de densidade, não é medida pelo grau dos nós (quantidade de arestas, métrica usualmente associada ao termo densidade), mas um reflexo da força das relações.

Assim, para efeito demonstrativo a métrica c*ohesiveness* foi estimada estaticamente, investigando o valor médio da intensidade das conexões dos subgrafos. Ou seja, a proporção relativa (razão) entre a força das conexões e a quantidade de arestas dos subgrafos induzidos ao longo do tempo. Desse modo, calculou-se um índice de densidade da força das relações, onde para cada ano acumula-se o subgrafo induzido a partir das conexões com *'voip',* somando-se o peso das arestas e dividindo pela quantidade de conexões (Equação 5). A Figura 48, inspirada na Figura 6, e a Tabela 4 exemplificam esse cálculo.

$$\begin{matrix} densidade\ da\ força \\ de\ conexão\ do \\ subgrafo \end{matrix} = \frac{soma\ dos\ pesos\ das\ conexões\ do\ subgrafo}{número\ de\ arestas\ do\ subgrafo} \qquad (5)$$

Figura 48 – Ilustração para o cálculo da densidade na intensidade das relações do subgrafo



Fonte: Elaborado pela autora inspirado em Chu *et al*. (2019)

Levando em conta a Figura 48, inspirada na ilustração apresentada no trabalho de Chu *et al*. (2019), suponha que o nó v_3 seja a palavra '*voip*'. Assim, no primeiro instante o subgrafo é formado pelos nós $v_1, v_2\ e\ v_3$, onde a densidade das conexões é estabelecida pela soma dos pesos (4) dividido pelo número de arestas (2), como pode ser visto na Tabela 4. Para o instante t_2, calcula-se a densidade a partir do subgrafo acumulado, somando o peso das arestas das relações presentes no subgrafo, resultando, em um índice de 2,5. Assim, sucessivamente, no acumulado após o instante t_3, tem-se a densidade igual a 3. É possível constatar, então, que ao decorrer do tempo existe um crescimento da intensidade das relações do subgrafo em torno do nó v_3.

No exemplo, fica estabelecido que os nós $v_1, v_2\ e\ v_3$ induzem um subgrafo no intervalo $[t_1, t_3]$ que maximiza a métrica *cohesiveness,* pois o grafo induzido por esse conjunto de nós nesse período apresenta a maior força de conexão média entre os nós. Como seria esperado, com o decorrer do tempo, as palavras que formam o contexto semântico de *'voip'* são conectadas entre elas, o que resulta em um subgrafo acumulado com uma maior intensidade, isto é, uma maior média de força nas conexões. Tal que, embora não tenha sido demonstrado operacionalmente a descoberta do subgrafo, percebe-se como ao decorrer dos anos a rede em torno da palavra nova se torna mais densa. Desta forma, o algoritmo TDF seria capaz de detectar esse crescimento de intensificação que ocorre no subgrafo num intervalo de tempo (*burst density*).

Tabela 4 – Demonstração da densidade da Figura 48.

	Instante t1	Instante t2	Acumulado (t1:t2)	Instante t3	Acumulado (t1:t3)
nº nós	3	2	3	3	3
nº arestas	2	1	2	2	3
soma dos pesos	4	1	5	4	9
densidade			2.5		3

Fonte: Elaborado pela autora

Estendendo esse raciocínio para o exemplo desta seção, a Figura 49 traz o cálculo da métrica. O código executa várias operações nos grafos representados como listas de adjacências, para cada ano considerado. Na sequência, acessa a lista de adjacência referente ao grafo obtido por meio da similaridade de cossenos como apresenta na Figura 44, acessa a lista de adjacência do subgrafo extraído em dois níveis como mostra a Figura 46 e, então, com o passar do tempo acumula os grafos, extrai os subgrafos e disponibiliza dados como o número

de nós, a quantidade de arestas e a soma das intensidades. A partir dessas informações estimou-se a densidade das forças de conexão buscando verificar o crescimento dessa intensidade nas relações.

Figura 49 – Obtenção dos subgrafos acumulados temporalmente e cálculo da métrica de densidade das forças.

```
#2002
adjacency_list = adjacency_list_2002
time = '2002'

graph = create_graph(adjacency_list)
print("Number of nodes - {}: {}".format(time, len(graph)))
print("Number of edges - {}: {}".format(time, graph.number_of_edges()))
sum_edges = get_sum_edges(adjacency_list)
print('Sum of the edge weights - {}: {}'.format(time, sum_edges))
nx.write_graphml_lxml(graph, path+time+".graphml")

adjacency_list_new = extract_subgraph(adjacency_list, token)
subgraph_new = create_graph(adjacency_list_new)
print("\nNumber of nodes - subgraph w/a: ", len(subgraph_new))
sum_edges = get_sum_edges(adjacency_list_new)
print("Number of edges - subgraph w/a: ", subgraph_new.number_of_edges())
print('Sum of the edge weights - subgraph w/a: ',sum_edges)

accumulate_relation(adjacency_list, accumulated_adjacency_list)
accumulated_graph = create_graph(accumulated_adjacency_list)
nx.write_graphml_lxml(accumulated_graph, path+time+"a.graphml")

print("\nNumber of nodes - accumulated: ", len(accumulated_graph))
print("Number of edges - accumulated: ", accumulated_graph.number_of_edges())
sum_edges = get_sum_edges(accumulated_adjacency_list)
print('Sum of the edge weights - accumulated: ',sum_edges)
```

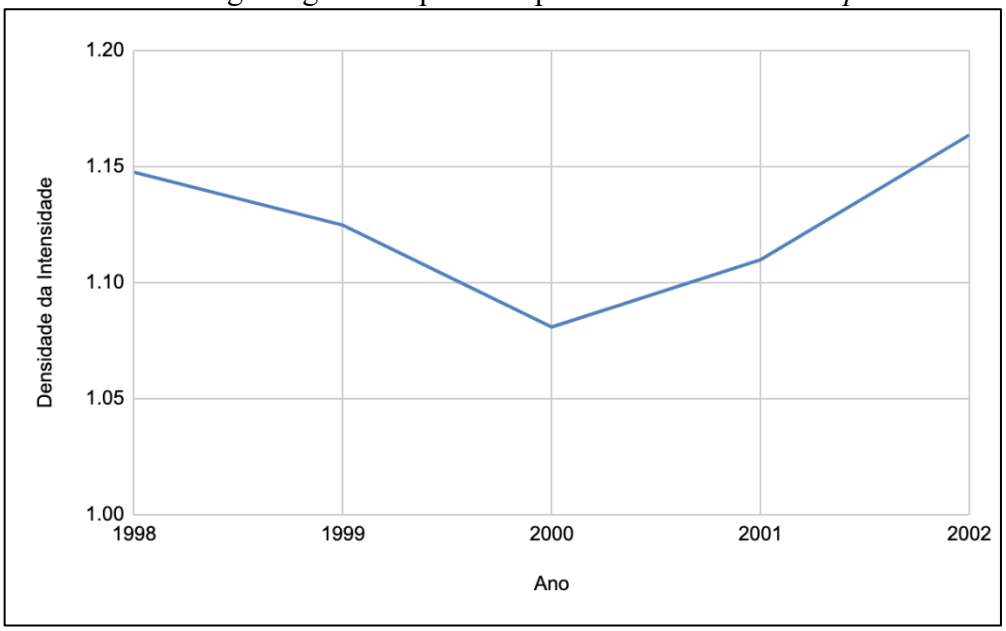
```
Number of nodes - 2002: 5744
Number of edges - 2002: 34164
Sum of the edge weights - 2002: 32114.272832930088

Number of nodes - accumulated:  14527
Number of edges - accumulated:  135115
Sum of the edge weights - accumulated:  135886.95536124706

Number of nodes - subgraph:  1372
Number of edges - subgraph:  3586
Sum of the edge weights - subgraph:  4172.805191397667
```

Fonte: Elaborado pela autora

No gráfico expresso na Figura 50 verifica-se que no decorrer dos anos houve um crescimento na intensificação das relações do subgrafo induzido a partir da palavra *'voip'*. Existe inicialmente um decréscimo seguido de uma subida, observações coerentes com o que pode-se constatar na evolução dos *embeddings* (Figura 42), quando a palavra *'voip'*, em 2000 (Figura 41), tem seu *embedding* começando a representar seu contexto adequado. É possível

notar uma mudança na direção da densidade das forças que começam a aumentar refletindo a intensificação nas relações, ou seja, as palavras desse contexto, desse subgrafo, estão relacionadas mais fortemente entre si. Relembrando, como visto na seção 2.3.2.1, que o subgrafo *burst* temporal apresenta crescimento mais rápido dessa densidade, não necessariamente o subgrafo mais denso. Sendo assim, pode-se inferir que a palavra *'voip'* estaria presente no conjunto de nós que induzem o subgrafo com maior *cohesiveness,* sendo identificada como um sinal latente por volta de 2001, quando se constata a mudança de direção da densidade das forças e intensificação na relação entre os nós do subgrafo.

Figura 50 – Gráfico da progressão temporal da taxa de densidade da intensidade das relações no subgrafo gerado a partir da palavra de interesse *'voip'*.



Fonte: Elaborado pela autora

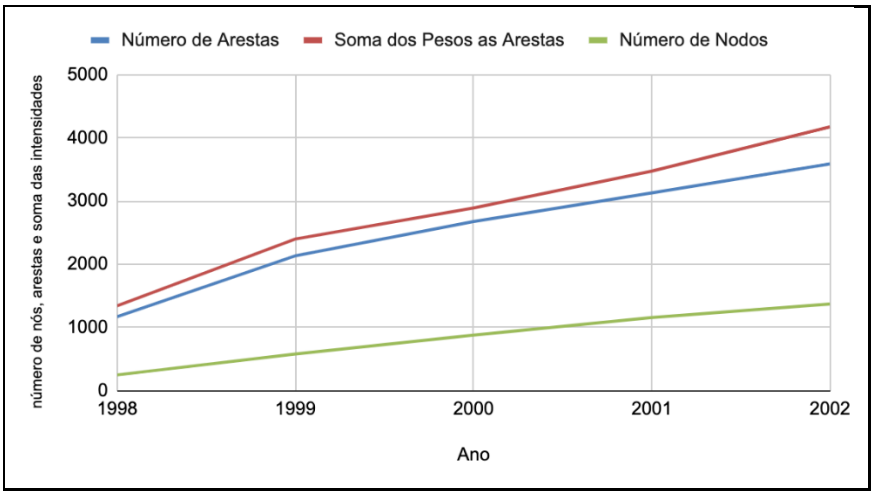
Pode-se notar também o descolamento do crescimento da soma da intensidade das relações em comparação com o crescimento da quantidade de nós da rede e aumento no número de conexões (arestas) (Figura 51 e Tabela 5).

Tabela 5 – Descritivo da quantidade de nós, arestas e soma das forças das conexões dos subgrafos

Ano	1998	1999	2000	2001	2002
nº nós	249	579	879	1.156	1.372
nº arestas	1.170	2.133	2.674	3.128	3.586
soma dos pesos	1.342,72	2.399,19	2.890,27	3.471,50	4.172,80

Fonte: Elaborado pela autora

Figura 51 – Gráfico do crescimento temporal no número de nós, arestas e soma da intensidade das conexões dos subgrafos.



Fonte: Elaborado pela autora

Como saída desta etapa tem-se então uma lista de sinais latentes, identificados a partir dos sinais novos pertencentes aos subgrafos de comportamento *burst*.

5.2.6 Etapa 6 - Identificar Sinais Fracos

Para identificar os sinais fracos é preciso estimar, dentre os sinais latentes, aqueles que apresentam um possível impacto tecnológico. Por esse motivo, é realizada uma consulta a partir da lista de sinais latentes (saída da etapa anterior) na lista de palavras extraídas da base de patentes, como mostra a Figura 52[40].

Para a lista de palavras da base de patentes, a partir dos documentos de pedidos de patentes, o texto é recuperado do campo '*title*' e '*abstract*' (semelhante à elaboração do *dataset* acadêmico na Etapa 1). Como pré-processamento, as palavras são padronizadas em caixa baixa, é utilizada uma *stoplist* para excluir palavras de pouca significância, tais como: preposições, conectivos, artigos, pronomes e outras. Ou seja, obtém-se uma lista de substantivos das patentes, palavras que representam seres, objetos, ações, lugares, etc.

Essa lista é cumulativa, a cada ano novas palavras são adicionadas. Essa decisão prevê o caso de ocorrência de palavras novas primeiro na fonte de informação tecnológica do que na fonte de informação científica, como no caso de publicações '*sleeping beauty*'. Nesse sentido, assim que a palavra for identificada na base acadêmica e como um sinal latente, ela será

[40] Palavras da lista tecnológica extraídas da primeira sentença do *abstract* da patente apresentada na Figura 30. Palavras da lista de sinais latentes, são palavras apresentadas como exemplos de palavras novas ao longo do texto e outras palavras também consideradas para simulação.

identificada como sinal fraco. Pois, mesmo que ocorra o caso no qual em anos posteriores a palavra não conste na base de patentes, ela já estaria adicionada na lista de palavras tecnológicas para quando ocorresse a consulta do modelo a partir da lista de palavras latentes.

Figura 52 – Demonstração da consulta da lista de sinais latentes na lista de palavras extraídas do título e resumo de pedidos de patentes considerando a palavra *'voip'* como sinal fraco.

```
#query the list of latent signals into the list of words from the patent base

#new words classified as latent signals due to their burst behaviour in the network
latent_signal_list = ['laser', 'smog','graphene', 'voip', 'metaverse', 'blockchain']

#nouns retrieved from the patent abstract
patent_words_list = ['system', 'voice', 'platform', 'voip', 'format', 'hardware', 'telephone', 'interface']

common_words = set(latent_signal_list) & set(patent_words_list)
print(common_words)
```

```
{'voip'}
```

Fonte: Elaborado pela autora

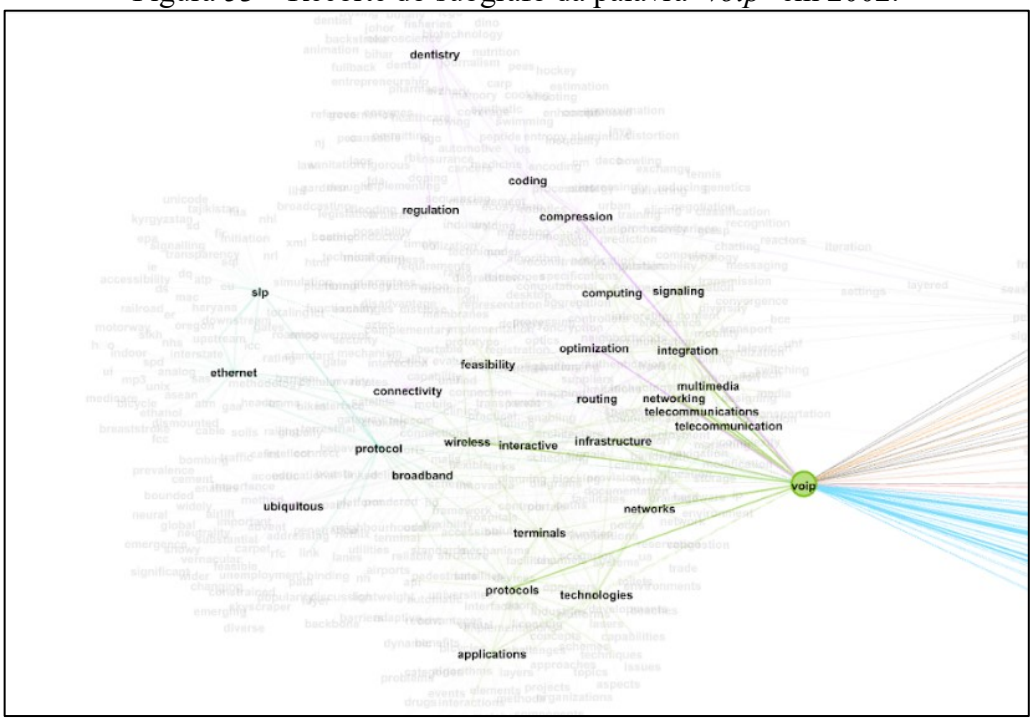
Nesta avaliação do modelo, o documento de patente do ano de 2001 apresenta a palavra *'voip'*, incluindo-a na lista. Assim, no início de 2002, a palavra '*voip*' seria identificada como sinal fraco, com a palavra '*voip*' identificada como sinal latente em 2001.

5.2.7 Etapa 7 - Apresentar Recomendações de Sinais Fracos para Tecnologias Emergentes

Uma vez encontrada uma sugestão de sinal fraco, deve-se retomar o subgrafo no qual este aparece a fim de contextualizar seu significado, oferecendo oportunidades de *insights* e novas possibilidades de pesquisas para que especialistas investigem o termo.

Na Figura 53 apresenta-se o recorte do subgrafo com a palavra '*voip*' apontando para seu contexto, onde pode-se ver a palavra *voip* associada a termos como '*telecommunication*', '*internet*', '*protocol*', entre outros, que descrevem o contexto de atuação de *voip* como uma possível nova tecnologia de telefonia via *internet*. Ou seja, a palavra nova identificada atuando como sinal fraco para uma potencial tecnologia emergente de um domínio tecnológico oculto.

Figura 53 – Recorte do subgrafo da palavra '*voip*' em 2002.



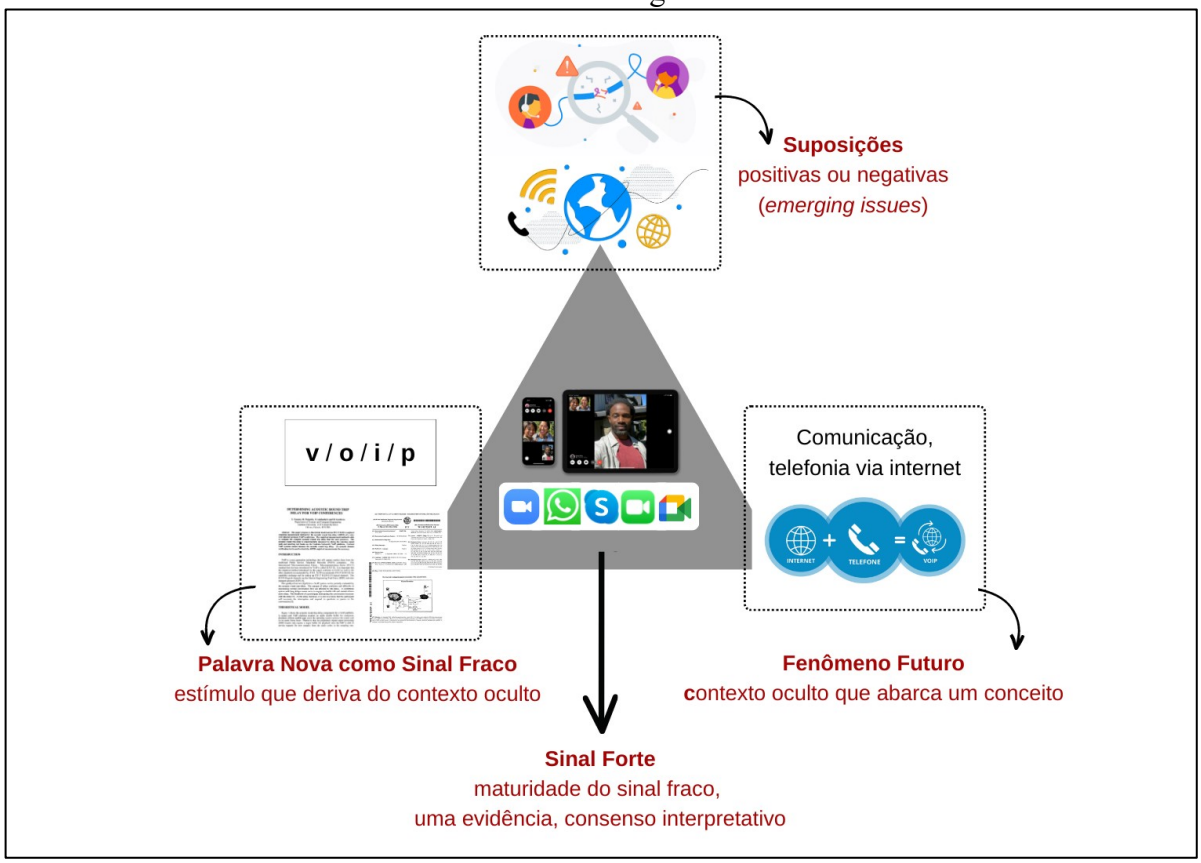
Fonte: Elaborado pela autora

5.3 CONCLUSÃO DO CAPÍTULO

Para finalizar, tomando como base que a avaliação de um artefato visa averiguar o quanto o modelo desenvolvido atende como solução para o problema identificado, considera-se que o objetivo foi atingido de forma bem-sucedida. Conjuntamente da apresentação do modelo pode-se acompanhar a fundamentação teórica que sustenta e justifica a elaboração e proposição de cada uma das etapas do modelo visando atender as deficiências encontradas na literatura. E, na aplicação operacional, apesar da execução limitada pela capacidade de processamento disponível, pode-se constatar indícios da viabilidade da proposta elaborada. Ressalta-se ainda, o rigor científico presente desde a concepção e condução da pesquisa que resultou no artefato apresentado.

Conclui-se este capítulo com o esquema representando a palavra *'voip'* como sinal fraco para a tecnologia emergente de chamadas de voz via *Internet.* Na Figura 54, é retratada a palavra *'voip'* como sinal fraco que remete ao contexto de telefonia via *internet,* ainda como uma tecnologia emergente, de potencial não explorado. Pela lógica apresentada pelo modelo, estima-se que '*voip*' seria sugerido como sinal fraco em 2001, sendo possível olhar para esse contexto tecnológico e seus potenciais impactos positivos e negativos com antecedência, permitindo determinar as dimensões estratégicas e o planejamento necessário objetivando garantir benefícios socioeconômicos.

Figura 54 – Representação da proposição da palavra nova ‘voip’ como sinal fraco, unificando os conceitos de Ansoff e do triângulo semiótico de Peirce.



Fonte: Elaborado pela autora (imagens obtidas no Google Imagens®)

6 CONSIDERAÇÕES FINAIS

A incerteza ocasionada pelo desenvolvimento acelerado enaltece o desejo pela antecipação da informação para o planejamento estratégico. A detecção de sinais fracos corresponde a essa demanda permitindo descobrir fragmentos de informação ainda incertos, mas que apontam para um potencial evento futuro, conferindo tempo hábil para tomadas de decisão antes que a informação se consolide. Particularmente, notou-se o interesse para acessar informações novas com antecedência a respeito do domínio tecnológico. Neste sentido, esta tese propôs um modelo para a coleta e identificação de sinais fracos de novidade para tecnologias emergentes, baseado em modelos de descoberta de conhecimento em texto a partir de publicações científicas e tecnológicas, como suporte para processos de planejamento estratégico.

O desafio de encontrar sinais fracos para o futuro está no presente. Esse processo de "adivinhar" o futuro depende muitas vezes de questões qualitativas e subjetivas. No entanto, necessita do apoio de técnicas quantitativas que impeçam que as decisões sejam tomadas sob algum viés cognitivo, promovendo um norte lógico ao raciocínio interpretativo. A adoção de uma visão Ansoffiana dos sinais fracos, na qual os fragmentos de informações existem independentes na capacidade de interpretação do especialista, implicam na possibilidade de implementar modelos suportados por KDT para identificá-los. Onde, entende-se um sinal fraco como uma ocorrência, como um dado, cuja interpretação é passível de gerar um aviso. Tal aviso indica que pode haver um acontecimento e decorrentes consequências, em termos de oportunidade ou ameaça. Assim, sinais fracos são tidos como fragmentos de informação a serem descobertos em meio a um grande volume de conteúdo produzido.

Contudo, as práticas quantitativas para obtenção de informações sobre tecnologias emergentes e também as abordagens para sinais fracos, de modo geral, apresentam análises de texto superficiais fundamentadas em técnicas estatísticas de frequência de palavras-chave, indicadores bibliométricos e redes de citações. Dessa forma, são abordagens que necessitam de mais tempo de observação, não suprindo a necessidade de antecedência de informação como oferece o viés de antecipação de novidade de sinais fracos para tecnologias emergentes a partir da identificação de novas palavras. Sendo assim, identificou-se a oportunidade de aprimoramento tanto do ponto de vista do nível de antecedência do sinal fraco, quanto das abordagens de mineração de texto explorando a semântica presente nos dados não estruturados.

.

Isto posto, consideram-se atendidos os objetivos específicos de identificar os fatores envolvidos na detecção de sinais fracos e analisar potenciais técnicas para sua detecção, bem como investigado os critérios que caracterizam o domínio tecnológico de interesse. Assim, pode-se estabelecer conceitualmente os sinais fracos para tecnologias emergentes e desenvolver um modelo, além de demonstrar sua capacidade de atender a necessidade identificada.

Como solução para o problema sugere-se a identificação de sinais fracos para tecnologias emergentes a partir da identificação de novas palavras como *proxies* capazes de apontar para um contexto de conhecimento tecnológico ainda oculto. Essa proposta tem como base conceitual o paralelo entre a teoria de sinais fracos de Ansoff com a teoria de Peirce, assim como proposto por Hiltunen (2008), e também a premissa histórica da necessidade de cunhar novas palavras a fim de compreender o mundo ao redor, fato que se destaca particularmente ao contexto de novas tecnologias, tendo em vista o processo de surgimento dessas nomenclaturas.

Verificou-se que termos tecnológicos são criados com o intuito de nomear novas descobertas e que o mercado atual (Ind. 4.0) está aquecido para o surgimento de novas tecnologias (tecnologias emergentes). Observou-se também que não é possível prever quando um novo termo será cunhado, porém é possível identificar seu surgimento a partir de seu primeiro uso. Nesse sentido, considerou-se a possibilidade de explorar o surgimento de novas palavras como um sinal fraco para tecnologias emergentes, compreendendo-as como o menor nível de abstração de uma informação, ou seja, um fragmento ainda vago capaz de sinalizar para um contexto oculto emergente estrategicamente relevante por seu potencial de impacto. Assim, por meio desse modelo de antecipação de conhecimento baseado na identificação de palavras novas como sinais fracos para tecnologias emergentes contribui-se para o planejamento estratégico, levando em consideração as incertezas do ambiente.

O modelo desenvolvido, a partir das diretrizes da DSRM, propôs uma abordagem considerando a semântica contextual das palavras, baseada na identificação de novas palavras e também recorreu à implementação de análises dinâmicas e temporais para detecção dos sinais fracos. Sustenta-se em três principais tarefas de mineração de texto: (i) identificação de palavras novas; (ii) monitoramento de seu rápido crescimento de interesse; e (iii) estimativa sobre o potencial de impacto. Tal proposição se faz factível em razão das novas potencialidades dos métodos e ferramentas de Processamento de Linguagem Natural e Ciência de Redes como recursos da Mineração de Texto.

Para a primeira tarefa, técnicas recentes de NLP explorando as potencialidades de LLMs foram implementadas na base de dados extraída de publicações científicas. A descoberta de novas palavras, é uma tarefa difícil, seguida do problema de obter seus *embeddings*, pois palavras raras, ou de vocabulário específico, costumam ser palavras OOV e ocorrem com baixa frequência, resultando em um contexto insuficiente para a geração de *embeddings* representativos. Contudo, pelo modelo se tratar de uma abordagem para detecção de sinais fracos, é intrínseca da teoria de sinais fracos a existência de um padrão evolutivo, permitindo que o modelo aperfeiçoe o *embedding* da palavra nova ao longo do monitoramento temporal dinâmico, sendo essa a tarefa subsequente a tarefa de identificação da palavra nova.

A segunda tarefa de mineração objetiva classificar as palavras novas (sinais novos) como sinais latentes, isto é, sinais intermediários que apresentam alguma relevância, mas ainda não são sinais fracos. Para tanto recorreu-se a representação das palavras em redes complexas temporais e a identificação de subgrafos com comportamento *burst,* refletindo a intensificação das relações em torno da palavra nova com as demais palavras que descrevem o seu contexto, indicando crescente interesse no tema.

Por fim, para identificação dos sinais fracos, na terceira tarefa de mineração, a relevância estratégica é determinada pela estimativa de impacto do sinal latente ao verificar sua projeção tecnológica conforme sua aparição em outras bases de conhecimento tecnológico. Assim, o resultado produzido é constituído por uma lista de palavras sugeridas como sinais fracos e seu contexto associado, de modo que a palavra como sinal fraco é um guia para esse contexto que contempla um nicho tecnológico emergente.

Ademais, considera-se que o objetivo da pesquisa foi cumprido com êxito. O conhecimento do tema teve avanços apresentando discussões a respeito da natureza dos sinais fracos e apresentou-se um novo modelo quantitativo para sua detecção a partir de uma abordagem inédita baseada em novas palavras. Todavia, limitações são inerentes ao processo, assim como proposições para desenvolvimentos futuros, apresentados na sequência.

6.1 LIMITAÇÕES

Inerente a qualquer modelo com base em dados, a primeira limitação do modelo a ser pontuada diz respeito à qualidade dos dados utilizados. Modelos de mineração de texto são

sensíveis à qualidade dos dados de entrada, podendo resultar em uma saída imprecisa. Esta questão é popularmente conhecida no meio da computação como "*trash-in/ trash-out*".

Segundo, ainda sobre o tipo de informação com a qual o modelo opera, reconhece-se que nem todo conhecimento está documentado, ou está em bases científicas, logo o modelo limita sua capacidade de antecipar conhecimento apenas a informações que estejam presentes em bases de dados textuais. Nessa direção, sua abordagem quantitativa não é apta a identificar sinais fracos resultantes de processos de ideação subjetivos e dependentes de filtragem cognitiva e enquadramento ideológico.

Pontua-se também limitações relativas à operacionalização do modelo. Enquanto teoricamente o modelo se sustenta, operacionalmente ainda existe uma barreira devido ao elevado custo de processamento de LLMs, mas principalmente de algoritmos de identificação de comunidades em grafos temporais, como a análise *burst* sugerida em uma das principais tarefas do modelo. Logo, conhecimentos operacionais de computação e programação e demais *softwares* são requeridos visando viabilizar a execução do modelo em sua plenitude.

Diante desse contexto, embora os resultados atingidos no desenvolvimento do método e no cenário de estudo, sejam promissores, para uma maior confiabilidade faz-se necessária a verificação do comportamento desta estratégia quando aplicada em contextos mais amplos, como na inclusão simultânea de várias palavras novas. A futura aplicação do método em outros cenários, poderá também esclarecer e determinar um número mínimo de ocorrências contextuais que uma palavra nova precisa ter, no *corpus* de treinamento, para produzir uma representação distribuída consistente. Demais desdobramentos futuros do modelo são abordados na seção seguinte.

Por fim, ressalta-se que a proposta elaborada nesta tese para detecção de sinais fracos para tecnologias emergentes é suscetível ao lapso temporal da aparição da palavra entre as duas fontes de informação a serem investigadas. No caso demostrado, o atraso entre a publicação científica e o pedido de patente. De forma similar, redes de citação tem sua fraqueza no tempo necessário entre a publicação e decorrentes citações. Mesmo assim, ainda que se tenha um atraso, presume-se que o modelo proposto oferece um caminho para descobrir com maior antecedência sinais fracos quando comparado com outros métodos e modelos que necessitam de frequência de ocorrência para identificação do sinal fraco.

6.2 ESTUDOS FUTUROS

São destacados aprimoramentos imediatos do modelo, bem como são elaboradas sugestões para estender a pesquisa.

De início, sugerem-se três pontos de investigações futuras para melhorar o modelo. Primeiramente, nota-se a necessidade de explorar e definir detalhes operacionais do modelo como: (i) o ponto de corte da similaridade entre os *embeddings* para estabelecer a rede de palavras; (ii) a quantidade *k*, de subgrafos, a serem identificadas pelo algoritmo TDF (*Top-k Density Burst Subgraphs Finding*); e (iii) o raio a ser considerado como distância para representação do subgrafo de domínio tecnológico do sinal fraco identificado. Para aperfeiçoar essas especificações julga-se relevante a aplicação do modelo em cenários variados visando observar como se comportam diferentes conjuntos de dados, auxiliando nessas decisões.

Segundo, vislumbra-se a possibilidade de estabelecer uma etapa posterior as etapas fornecidas no modelo instituindo estratégias de hierarquização dos sinais recomendados, por exemplo, considerando o intervalo de tempo (datas) entre sua identificação como latente e sinal fraco, ou de palavra nova ao sinal latente. Porém, novamente, observar como se comportaria o modelo a partir de casos com comportamentos variados promoveria clareza sobre quais estratégias seriam mais adequadas.

Terceira e última melhoria imediata de desenvolvimento do modelo relaciona-se à expansão da etapa de identificação de sinais novos, não apenas para palavras novas, mas também para palavras em novos contextos, como no caso dos acrônimos SPIDER, SHRIMP, etc., assim como discutido na seção 2.4. O modelo foi concebido tendo isso em mente, no entanto, ainda se faz necessário o aprofundamento nos estudos em NLP e identificação de mudanças semânticas (*semantic shift*). Conjectura-se ainda, que com essas etapas operacionais realizadas, possivelmente, detalhes não considerados até o momento no modelo proposto poderão surgir e serão refinados.

De desdobramentos operacionais, sugere-se a investigação de padrões que podem surgir ao se recuperar informações como, data e autoria das publicações associadas aos sinais fracos identificados. Também, sugere-se uma investigação comparativa com demais estudos averiguando a taxa de antecedência que o modelo aqui proposto apresentaria frente aos sinais identificados pelos estudos como JRC (EULAERTS *et al.*, 2019), KISTI (YANG *et al*., 2022), Zhu e Du (2023).

Como sugestões mais amplas para expandir a pesquisa, aponta-se a relevância de se considerar a dinâmica do comportamento de evolução dos sinais fracos baseados em palavras novas. Sugere-se então, um aprofundamento na investigação sobre como esses sinais baseados em palavras novas evoluem até sinais fortes e possivelmente agregar essa funcionalidade de monitoramento ao modelo de identificação.

Outra sugestão decorre da análise acerca da capacidade do modelo em executar suas tarefas em tempo real. O modelo foi exemplificado considerando o processamento anual das informações coletadas dos bancos de dados, entretanto, presume-se que o modelo seria capaz de comportar o processamento de fluxo de informações em tempo real. Contudo, um aprofundamento no estudo da viabilidade de processamento computacional se faz necessário.

Ainda, vislumbra-se sobre a adaptabilidade do modelo em outros contextos, por exemplo, investigar termos que surgem na *internet* e redes sociais como 'quarantini'[41], um neologismo relacionado a bebida Martini e quarentena ocasionada pela pandemia em 2020, como possível parte do contexto de aumento no consumo de bebida alcoólica[42]. Além disso, especula-se sobre a aplicação de fontes variadas de dados além do uso de dados textuais.

Ao longo da pesquisa surgiram ainda questionamentos para investigações futuras que compreendem desafios mais abrangentes para a área. Como a elaboração de modelos que integrem mais etapas do processo de análise de sinais fracos. Ou seja, investigar meios de automatizar ou semi-automatizar a primeira e segunda etapa dos sinais fracos ((1) coleta e identificação e (2) análise e diagnóstico), visando elaborar um produto final com todo o processo de sinais fracos para tecnologias emergentes que não seja dependente de *insights* de especialistas e filtragens manuais.

Vislumbra-se também sobre como estratégias poderiam abranger a análise sendo capazes de lidar com sinais fracos compostos por mais de um domínio. Reflete-se sobre como se poderia realizar tal feito de agrupar dois ou mais tipos de sinais fracos, incluindo sinais fracos políticos, econômicos, tecnológicos, sociais ou culturais, quais vantagens e dificuldades existiriam em analisar um sinal fraco composto ao invés de analisar um único sinal fraco.

---

[41] Reportagem do jornal "New York Post" sobre o drink 'quarentini': https://nypost.com/2020/03/17/heres-how-to-make-a-perfect-quarantini-while-stuck-at-home/

[42] Reportagem da OECD sobre o consumo de bebida alcoólica na pandemia: https://www.oecd.org/coronavirus/policy-responses/the-effect-of-covid-19-on-alcohol-consumption-and-policy-responses-to-prevent-harmful-alcohol-consumption-53890024/

Por fim, sugere-se um aprofundamento teórico investigando a reflexão levantada a partir da comparação proposta de sinais fracos enquanto palavras, frente à abordagem de signos futuros de Hiltunen. Indaga-se a respeito da correlação entre a natureza distinta dos sinais fracos (como observação e como interpretação) com o campo de *Foresight*, relacionando a visão ansoffiana com a antecipação de futuros possíveis, e a visão construtivista com a antecipação de futuros desejáveis. Sendo necessário examinar detalhadamente a teoria para sustentar essas afirmações propostas.

REFERÊNCIAS

AALTONEN, Mika; SANDERS, T. Irene. Identifying Systems' New Initial Conditions as Influence Points for the Future. **Foresight**, v. 8, n. 3, p. 28–35, mai. 2006. https://doi.org/10.1108/14636680610668054

AGUILAR, F. J.. **Scanning the business environment**. New York: Macmillan, 1967.

ADIL, Bouktaib; ABDELHADI, Fennan. A Framework for Weak Signal Detection in Competitive Intelligence Using Semantic Clustering Algorithms. **International Journal of Advanced Computer Science and Applications**, v. 12, n. 12, 2021. https://doi.org/10.14569/IJACSA.2021.0121271

AHLQVIST, Toni; UOTILA, Tuomo. Contextualising weak signals: towards a relational theory of futures knowledge. **Futures**, [S.L.], v. 119, p. 102543, maio 2020. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2020.102543

AHMAD, K. Neologisms, Nonces and Word Formation. **In Proceedings of the 9th EURALEX Int. Congress**, Munich: Universitat Stuttgart (ISBN 3-00-006574-1), v.2, p. 711-730, aug. 2000.

ALLAHYARI M.; POURIYEH, S.; ASSEFI, M.; SAFAEI, S.; TRIPPE, E. D.; GUTIERREZ, J. B.; KOCHUT, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. **In Proceedings of KDD Bigdas**, Halifax, Canada, p. 1-13, aug, 2017.

ALLAN, James; CARBONELL, Jaime G.; DODDINGTON, George; YAMRON, Jonathan; YANG, Yiming. Topic detection and tracking pilot study final report. **Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop**, fev. 1998.

ALMEIDA, Fernando C. de; LESCA, Humbert. Collective Intelligence Process to Interpret Weak Signals and Early Warnings. **Journal of Intelligence Studies in Business**, v. 9, n. 2, nov. 2019. https://doi.org/10.37380/jisib.v9i2.466

ANSOFF, H. Igor. Managing Strategic Surprise by Response to Weak Signals. **California Management Review**, [S.L.], v. 18, n. 2, p. 21-33, dez. 1975. SAGE Publications. http://dx.doi.org/10.2307/41164635

ANSOFF, H. Igor. Competitive strategy analysis on the personal computer. **Journal of Business Strategy**, v. 6, n. 3, p. 28-36, 1986. https://doi.org/10.1108/eb039117

ANSOFF, I; MCDONNELL, E.E. **Implanting Strategic Management**, 2ed., Prentice/Hall International Inc., New York, 1990.

AMANATIDOU, Effie; BUTTER, Maurits; CARABIAS, Vicente; KONNOLA, Totti; LEIS, Miriam; SARITAS, Ozcan; SCHAPER-RINKEL, Petra; RIJ, Victor van. On Concepts and Methods in Horizon Scanning: Lessons from Initiating Policy Dialogues on Emerging Issues. **Science and Public Policy**, v. 39, n. 2, p. 208–21, mar. 2012. https://doi.org/10.1093/scipol/scs017

AMANATIDOU, Effie; CARABIAS-BARCELO, Vicente; LEIS, Miriam; SARITAS, Ozcan; SCHAPER-RINKEL, PETRA; SCHOONHOVEN, Bas van; RIJ, Victor van; WARRINGTON,

Brian. Scanning for Emerging Science and Technology Issues. **European Foresight Plataform: Foresight Brief No. 197**, jul. 2011.

AMPLAYO, Reinald Kim; HONG, SuLyn; SONG, Min. Network-Based Approach to Detect Novelty of Scholarly Literature. **Information Sciences**, v. 422, p. 542–57, jan. 2018. https://doi.org/10.1016/j.ins.2017.09.037

ARAUJO, Miguel, *et al*. Discovery of 'Comet' Communities in Temporal and Labeled Graphs COM^2. **Knowledge and Information Systems**, v. 46, n. 3, p. 657–77, Mar. 2016. https://doi.org/10.1007/s10115-015-0847-2.

ÁVILA-ROBINSON, Alfonso; MIYAZAKI, Kumiko. Dynamics of Scientific Knowledge Bases as Proxies for Discerning Technological Emergence — The Case of MEMS/NEMS Technologies. **Technological Forecasting and Social Change**, v. 80, n. 6, p. 1071–84, jul. 2013. https://doi.org/10.1016/j.techfore.2012.07.012

BA, Zhichao; LIANG, Zhentao. A Novel Approach to Measuring Science-Technology Linkage: From the Perspective of Knowledge Network Coupling. **Journal of Informetrics**, v. 15, n. 3, p. 101167, ago. 2021. https://doi.org/10.1016/j.joi.2021.101167

BARABASI, Albert-László. Network Science. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 371, n. 1987, p. 20120375, mar. 2013. https://doi.org/10.1098/rsta.2012.0375

BARABASI, Albert-László; BONABEAU, Eric. Scale-Free Networks. **Scientific American**, v. 288, n. 5, p. 60–69, mai. 2003. https://doi.org/10.1038/scientificamerican0503-60

BARABASI, Albert-László; ALBERT, Réka. Emergence of Scaling in Random Networks. **Science**, v. 286, n. 5439, p. 509–12, out. 1999. https://doi.org/10.1126/science.286.5439.509

BARNES, James H.. Cognitive biases and their impact on strategic planning. **Strategic Management Journal**, [S.L.], v. 5, n. 2, p. 129-137, abr. 1984. Wiley. http://dx.doi.org/10.1002/smj.4250050204

BARAM-TSABARI; A.; WOLFSON, O.; YOSEF, R.; CHAPNIK, N.; BRILL, A.; SEGEV, E. Jargon use in Public Understanding of Science papers over three decades. Public Understanding of Science, [S.L.], v. 29, n. 6, p. 644-654, 2020. SAGE. http://dx.doi.org/10.1177/0963662520940501

BASS, Scott D.; KURGAN, Lukasz A.. Discovery of Factors Influencing Patent Value Based on Machine Learning in Patents in the Field of Nanotechnology. **Scientometrics**, v. 82, n. 2, p. 217–41, fev. 2010. https://doi.org/10.1007/s11192-009-0008-z

BELTAGY, Iz.; LO, Kyle.; COHAN, Arman.. SciBERT: Pretrained Contextualized *Embeddings* for Scientific Text. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, [S.L.], 2019. http://dx.doi.org/10.18653/v1/D19-1371

BENNETT, Nathan; LEMOINE, G. James. What a difference a word makes: understanding threats to performance in a vuca world. **Business Horizons**, [S.L.], v. 57, n. 3, p. 311-317, maio 2014. Elsevier BV. http://dx.doi.org/10.1016/j.bushor.2014.01.001

BEREZNOY, Alexey. Corporate Foresight in Multinational Business Strategies. **Foresight And Sti Governance**, [S.L.], v. 11, n. 1, p. 9-22, 28 mar. 2017. National Research University, Higher School of Economics (HSE). http://dx.doi.org/10.17323/2500-2597.2017.1.9.22

BERNARD, Gilles; LEBBOSS, Georges. Methods for Word Encoding: A Survey. **Proceedings International Conference on Engineering and Technology** (ICET-IEEE), p. 1–6, 2017. https://doi.org/10.1109/ICEngTechnol.2017.8308139

BETZ, Ulrich A.K.; BETZ, Frederick; KIM, Rachel; MONKS, Brendan; PHILLIPS, Fred. Surveying the future of science, technology and business – A 35 year perspective. **Technological Forecasting And Social Change**, [S.L.], v. 144, p. 137-147, jul. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2019.04.005

BIDERMAN, M.T.C. **As ciências do léxico**. In OLIVEIRA de, A. M. P. P.; ISQUERDO, A. N. (orgs.). Às ciências do léxico: lexicologia, lexicografia, terminologia, 2ªed. Campo Grande: Ed. UFMS/INEP, 2001.

BILDOSOLA, Iñaki, GONZALEZ, Pilar; MORAL, Paz. An Approach for Modelling and Forecasting Research Activity Related to an Emerging Technology. **Scientometrics**, v. 112, n. 1, p. 557–72, jul. 2017. https://doi.org/10.1007/s11192-017-2381-3

BILDOSOLA, Iñaki; RIO-BELVER, Rosa María; GARECHANA, Gaizka; CILLERUELO, Ernesto. TeknoRoadmap, an Approach for Depicting Emerging Technologies. **Technological Forecasting and Social Change**, v. 117, p. 25–37, abr. 2017. https://doi.org/10.1016/j.techfore.2017.01.015

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex Networks: Structure and Dynamics. **Physics Reports**, v. 424, n. 4–5, p. 175–308, fev. 2006. https://doi.org/10.1016/j.physrep.2005.10.009

BONACCORSI, Andrea; APREDA, Riccardo; FANTONI, Gualtiero. Expert biases in technology foresight. Why they are a problem and how to mitigate them. **Technological Forecasting And Social Change**, [S.L.], v. 151, p. 119855, fev. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2019.119855

BORNMANN, Lutz; MUTZ, Rüdiger. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. **Journal Of The Association For Information Science And Technology**, [S.L.], v. 66, n. 11, p. 2215-2222, 29 abr. 2015. Wiley. http://dx.doi.org/10.1002/asi.23329

BOJANOWSKI, Pior; GRAVE, Edouard; JOULIN, Armand; MIKOLOV, Tomas. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, v.5, p.135–146, 2017. http://dx.doi.org/10.1162/tacl_a_00051

BREINER, Sibylle; CUHLS, Kerstin; GRUPP, Hariolf. Technology Foresight Using a Delphi Approach: A Japanese-German Co-Operation. **R&D Management**, v. 24, n. 2, p. 141–53, abr. 1994. https://doi.org/10.1111/j.1467-9310.1994.tb00866.x

BROWN, Tom, *et al*. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p.1877-1901, 2020. arXiv:2005.14165

BUCKLAND, Michael K. What is a “document”?. **Journal of the American society for information science**, v. 48, n. 9, p. 804-809, 1997. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<804::AID-ASI5>3.0.CO;2-V

BUNGE, M. Philosophy of science and technology: parte II: life science, social science and technology. **Treatise on basic philosophy**, Vol. 7. Dordrecht: Reidel, 1985.

BURMAOGLU, Serhat; SARTENAER, Olivier; PORTER, Alan; LI, Munan. Analysing the theoretical roots of technology emergence: an evolutionary perspective. **Scientometrics**, [S.L.], v. 119, n. 1, p. 97-118, 18 fev. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11192-019-03033-y

BURMAOGLU, Serhat; SARTENAER, Olivier; PORTER, Alan. Conceptual definition of technology emergence: a long journey from philosophy of science to science policy. **Technology In Society**, [S.L.], v. 59, p. 101126, nov. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.techsoc.2019.04.002

CAGNIN, Cristiano; HAVAS, Attila; SARITAS, Ozcan. Future-Oriented Technology Analysis: Its Potential to Address Disruptive Transformations. **Technological Forecasting and Social Change**, v. 80, n. 3, p. 379–85, mar. 2013. https://doi.org/10.1016/j.techfore.2012.10.001

CARLEY, Stephen F.; NEWMAN, Nils C.; PORTER, Alan L.; GARNER, Jon G. An Indicator of Technical Emergence. **Scientometrics**, v. 115, n. 1, p. 35–49, abr. 2018. https://doi.org/10.1007/s11192-018-2654-5

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS (CGEE). Metodologia para identificação de sinais fracos e monitoramento de tendências globais em CT&I. Brasília, 2016

CHAKRABARTI, A. A course for teaching design research methodology. **Artificial Intelligence for Engineering Design, Analysis and Manufacturing**, v. 24, p. 317-334, 2010. http://dx.doi.org/10.1017/S0890060410000223

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly Detection: A Survey. **ACM Computing Surveys**, v. 41, n. 3, p. 1–58, jul. 2009. https://doi.org/10.1145/1541880.1541882

CHEN, Chaomei. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. **Journal of the American Society for Information Science and Technology**, v. 57, n. 3, p. 359–77, fev. 2006. https://doi.org/10.1002/asi.20317

CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big Data: a survey. **Mobile Networks And Applications**, [S.L.], v. 19, n. 2, p. 171-209, 22 jan. 2014. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11036-013-0489-0

CHIARELLO, Filippo; TRIVELLI, Leonello; BONACCORSI, Andrea; FANTONI, Gualtiero. Extracting and Mapping Industry 4.0 Technologies Using Wikipedia. **Computers in Industry**, v. 100, p. 244–57, set. 2018. https://doi.org/10.1016/j.compind.2018.04.006

CHOWDHURY, G. Natural language processing, **Annual Review of Information Science and Technology**, v. 37, pp. 51-89, 2003. ISSN 0066-4200

CHU, Lingyang; ZHANG, Yanyan; YANG, Yu; WANG, Lanjun; PEI, Jian. Online density bursting subgraph detection from temporal graphs. **In Proceedings Of The Vldb Endowment**, [S.L.], v. 12, n. 13, p. 2353-2365, set. 2019. VLDB Endowment. http://dx.doi.org/10.14778/3358701.3358704

COATS, Daniel R. National Intelligence Strategy of the United States of America 2019. United States Office of the Director of National Intelligence, 2019. Disponível em: https://www.dni.gov/files/ODNI/documents/National_Intelligence_Strategy_2019.pdf

COATES, Vary; FAROOQUE, Mahmud; KLAVANS, Richard; LAPID, Koty; LINSTONE, Harold A.; PISTORIUS, Carl; PORTER, Alan L.. On the Future of Technological Forecasting. **Technological Forecasting and Social Change**, v. 67, n. 1, p. 1–17, mai. 2001. https://doi.org/10.1016/S0040-1625(00)00122-0

COATES, Joseph; DURANCE, Philippe; GODET, Michel. Strategic Foresight Issue: Introduction. **Technological Forecasting and Social Change**, v. 77, n. 9, p. 1423–25, nov. 2010. https://doi.org/10.1016/j.techfore.2010.08.001

COZZENS, Susan; GATCHAIR, Sonia; KANG, Jongseok; KIM, Kyung-Sup; LEE, Hyuck Jai; ORDÓÑEZ, Gonzalo; PORTER, Alan. Emerging technologies: quantitative identification and measurement. **Technology Analysis & Strategic Management**, [S.L.], v. 22, n. 3, p. 361-376, abr. 2010. Informa UK Limited. http://dx.doi.org/10.1080/09537321003647396

CRESWELL, John W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**, 3. ed., Porto Alegre: Artmed, 2010.

CUHLS, Kerstin; BEYAER-KUTZNER, Amina; GANZ, Walter; WARNKE, Philine. The Methodology Combination of a National Foresight Process in Germany. **Technological Forecasting and Social Change**, v. 76, n. 9, p. 1187–97, Nov. 2009. https://doi.org/10.1016/j.techfore.2009.07.010

CUPANI, A. **Filosofia da tecnologia: um convite**, 3. ed, Florianopolis: UFSC, 2016.

DAI, Jie; LI, Yuan; FAN, Xiaolin; SUN, Jing; ZHAO, Yuhai. Finding Early Bursting Cohesive Subgraphs in Large Temporal Networks. **IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation** (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), p. 264–71, 2021. https://doi.org/10.1109/SWC50871.2021.00044.

DAIM, T.U.; OLIVER, T. Implementing Technology Roadmap Process in the Energy Services Sector: A Case Study of a Government Agency. **Technological Forecasting and Social Change**, v. 75, p. 687-720, 2008. http://dx.doi.org/10.1016/j.techfore.2007.04.006

DAIM, Tugrul U.; RUEDA, Guillermo; MARTIN, Hilary; GERDSRI, Pisek. Forecasting Emerging Technologies: Use of Bibliometrics and Patent Analysis. **Technological Forecasting and Social Change**, v. 73, n. 8, p. 981–1012, out. 2006. https://doi.org/10.1016/j.techfore.2006.04.004

DARGAN, Shaveta; KUMAR, Munish; AYYAGARI, Maruthi Rohit; KUMAR, Gulshan. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning.

Archives of Computational Methods in Engineering, v. 27, n. 4, p. 1071–92, set. 2020. https://doi.org/10.1007/s11831-019-09344-w

DAY, George S.; SHOEMAKER, Paul J. H.. A different game, **Wharton on Managing Emerging Technologies**. John Wiley and Sons Inc., Hoboken, New Jersey, p. 1–23, 2000.

DAY, George S.; SCHOEMAKER, Paul J.H. Peripheral Vision: Detecting the Weak Signals that will Make or Break Your Company, Harvard Business School Press, Boston, 2006.

DERNIS, Hélène; SQUICCIARINI, Mariagrazia; PINHO, Roberto de. Detecting the emergence of technologies and the evolution and co-development trajectories in science (DETECTS): a 'burst' analysis-based approach. **The Journal Of Technology Transfer**, [S.L.], v. 41, n. 5, p. 930-960, 24 out. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10961-015-9449-0

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)**, [S.L], 2019. http://dx.doi.org/10.18653/v1/N19-1423

DONDI, Riccardo; HOSSEINZADEH, Mohammad Mehdi. Dense Sub-Networks Discovery in Temporal Networks. **SN Computer Science**, v. 2, n. 3, p. 158, maio 2021. https://doi.org/10.1007/s42979-021-00593-w.

DOTSIKA, Fefie; WATKINS, Andrew. Identifying Potentially Disruptive Trends by Means of Keyword Network Analysis. **Technological Forecasting and Social Change**, v. 119, p. 114–27, jun. 2017. https://doi.org/10.1016/j.techfore.2017.03.020

DRESCH, A.; LACERDA, D.P.; ANTUNES, J.A.V. **Design Science Research**. Springer, 2015. https://doi.org/10.1007/978-3-319-07374-3_4

DREYER, Daniel R.; RUOFF, Rodney S.; BIELAWSKI, Christopher W.. From Conception to Realization: An Historial Account of Graphene and Some Perspectives for Its Future. **Angewandte Chemie International Edition**, v. 49, n. 49, p. 9336–44, dez. 2010. https://doi.org/10.1002/anie.201003024

DU, Jian, WU, Yishan. A Parameter-Free Index for Identifying under-Cited Sleeping Beauties in Science. **Scientometrics**, v. 116, n. 2, p. 959–71, ago. 2018. https://doi.org/10.1007/s11192-018-2780-0

EBADI, Ashkan, AUGER, Alain; GAUTHIER, Yvan. Detecting Emerging Technologies and Their Evolution Using Deep Learning and Weak Signal Analysis. **Journal of Informetrics**, v. 16, n. 4, p. 101344, nov. 2022. https://doi.org/10.1016/j.joi.2022.101344.

EEROLA, A.; MILES, I.. Methods and tools contributing to FTA: a knowledge-based perspective. **Futures**, [S.L.], v. 43, n. 3, p. 265-278, abr. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2010.11.005

EFIMENKO, Irina V.; KHOROSHEVSKY, Vladimir F. Peaks, Slopes, Canyons and Plateaus: Identifying Technology Trends Throughout the Life Cycle. **International Journal of Innovation and Technology Management**, v. 14, n. 02, p. 1740012, abr. 2017. https://doi.org/10.1142/S0219877017400120

EISENHARDT, Kathleen M; BROWN, Shona L. Competing on the Edge: strategy as structured chaos. **Long Range Planning**, [S.L.], v. 31, n. 5, p. 786-789, out. 1998. Elsevier BV. http://dx.doi.org/10.1016/s0024-6301(98)00092-2

ENA, Oleg; MIKOVA, Nadezhda; SARITAS, Ozcan; SOKOLOVA, Anna. A Methodology for Technology Trend Monitoring: The Case of Semantic Technologies. **Scientometrics**, v. 108, n. 3, p. 1013–41, set. 2016. https://doi.org/10.1007/s11192-016-2024-0

ERDÖS, P.; RÉNYI, A. On random graphs, I. **Publicationes Mathematicae** (Debrecen), v. 6, p. 290–297, 1959.

ERNST, Holger. Patent Information for Strategic Technology Management. **World Patent Information**, v. 25, n. 3, p. 233–42, set. 2003. https://doi.org/10.1016/S0172-2190(03)00077-2

ERZURUMLU, S. Sinan. 4Cs of Innovation: A Conceptual Framework for Evaluating Innovation Strategy. **IEEE Engineering Management Review**, v. 45, n. 3, p. 42–53, 2017. https://doi.org/10.1109/EMR.2017.2734321

ERZURUMLU, S. Sinan; PACHAMANOVA, Dessislava. Topic Modeling and Technology Forecasting for Assessing the Commercial Viability of Healthcare Innovations. **Technological Forecasting and Social Change**, v. 156, p. 120041, jul. 2020. https://doi.org/10.1016/j.techfore.2020.120041

EULAERTS; O; JOANNY, G.; GIRALDI, J.; FRAGKISKOS, S.; PERANI, S.. Weak signals in Science and Technologies - Report: : Technologies at a Very Early Stage of Development That Could Impact the Future, EUR 29900 EN. **Publications Office of the European Union**, Luxembourg, 2019. ISBN 978-92-76-12387-3. https://data.europa.eu/doi/10.2760/50544

EULAERTS, O.; JOANNY, G.; GIRALDI, J.; FRAGKISKOS, S.; BREMBILLA, S.; ROSSI, D.; NICULA, G.; PERANI, S.. Weak signals in Science and Technologies - Weak signals in 2020, EUR 30714 EN, **Publications Office of the European Union**, Luxembourg, 2021, ISBN 978-92-76-37956-0. https://doi.org/10.2760/453777

EULAERTS, O.; JOANNY, G.; FRAGKISKOS, S.; GRABOWSKA, M.; BREMBILLA, S.; ROSSI, D.; NICULA, G.; PERANI, S.. Weak signals in Science and Technologies in 2021, EUR 31171 EN, **Publications Office of the European Union, Luxembourg**, 2022, ISBN 978-92-76-55597-1. https://doi.org/10.2760/700257

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p.37, 1996.

FELDMAN, Ronen; DAGAN, Ido. Knowledge Discovery in Textual Databases (KDT). **In Proceedings KDD**, v. 95, p. 112-117, 1995.

FLEMING, Lee; SORENSON, Olav. Science as a Map in Technological Search. **Strategic Management Journal**, v. 25, n. 89, p. 909–28, ago. 2004. https://doi.org/10.1002/smj.384

FRADKOV, Alexander L.. Early History of Machine Learning. **Ifac-Papersonline**, [S.L.], v. 53, n. 2, p. 1385-1390, 2020. Elsevier BV. http://dx.doi.org/10.1016/j.ifacol.2020.12.1888

FURUKAWA, Takao; MORI, Kaoru; ARINO, Kazuma; HAYASHI, Kazuhirio; SHIRAKAWA, Nobuyuki. Identifying the Evolutionary Process of Emerging Technologies: A Chronological Network Analysis of World Wide Web Conference Sessions. **Technological Forecasting and Social Change**, v. 91, p. 280–94, fev. 2015. https://doi.org/10.1016/j.techfore.2014.03.013

GALATI, Francesco; BIGLIARDI, Barbara. Industry 4.0: Emerging Themes and Future Research Avenues Using a Text Mining Approach. **Computers in Industry**, v. 109, p. 100–13, ago. 2019. https://doi.org/10.1016/j.compind.2019.04.018

GAO, Xubo; ZHENG, Qiusheng; VEGA-OLIVEIROS, Didier A.; ANGHINONI, Leandro; ZHAO, Liang. Temporal Network Pattern Identification by Community Modelling. **Scientific Reports**, v. 10, n. 1, p. 240, jan. 2020. https://doi.org/10.1038/s41598-019-57123-1.

GARCIA-NUNES, Pedro Ivo; RODRIGUES, Pedro Artico; OLIVEIRA, Kaulitz Guimarães; SILVA, Ana Estela Antunes da. A Computational Tool for Weak Signals Classification – Detecting Threats and Opportunities on Politics in the Cases of the United States and Brazilian Presidential Elections. **Futures**, v. 123, p. 102607, out. 2020. https://doi.org/10.1016/j.futures.2020.102607

GEORGHIOU, Luke. The UK Technology Foresight Programme. **Futures**, v. 28, n. 4, p. 359–77, may 1996. https://doi.org/10.1016/0016-3287(96)00013-4

GIBSON, Elizabeth; DAIM, Tugrul; GARCES, Edwin; DABIC, Marina. A Bibliometric Analysis to Identify Leading and Emerging Methods. **Foresight and STI Governance**, v. 12, n 1, p. 6–24, mar. 2018. https://doi.org/10.17323/2500- 2597.2018.1.6.24

GIRVAN, M.; NEWMAN, M. E. J.. Community Structure in Social and Biological Networks. **Proceedings of the National Academy of Sciences**, v. 99, n. 12, p. 7821–26, jun. 2002. https://doi.org/10.1073/pnas.122653799

GODET, Michel. The Art of Scenarios and Strategic Planning. **Technological Forecasting and Social Change**, v. 65, n. 1, p. 3–22, set. 2000. https://doi.org/10.1016/S0040-1625(99)00120-1

GORDON, Adam Vigdor; RAMIC, Mirza; ROHRBECK, René; SPANIOL, Matthew J.. 50 Years of corporate and organizational foresight: looking back and going forward. **Technological Forecasting And Social Change**, [S.L.], v. 154, p. 119966, maio 2020. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2020.119966

GOMILA, J. M.V.; RAMIREZ, M. A. A.; TING, M.; PORTER, A. L. Combining Tech Mining and Semantic TRIZ for Technology Assessment: Dye-Sensitized Solar Cell as a Case. Technological Forecasting and Social Change, v. 169, p. 120826, ago. 2021. https://doi.org/10.1016/j.techfore.2021.120826

GRIBKOVSKAIA, I.; HALSKAU, O.; LAPORTE, G. General Solutions to the Single Vehicle Routing Problem with Pickups and Deliveries. **European Journal of Operational Research**, v. 180, n. 2, p. 568–84, jul. 2007. https://doi.org/10.1016/j.ejor.2006.05.009.006

GRIOL-BARRES, Israel; MILLA, Sergio; MILLET, Jose. Improving Strategic Decision Making by the Detection of Weak Signals in Heterogeneous Documents by Text Mining

Techniques. **AI Communications**, v. 32, n. 5-6, p. 347-360, jan. 2019. http://doi.org/10.3233/AIC-190625

GRIOL-BARRES, Israel; MILLA, Sergio; CEBRIAN, Antonio ; FAN, Huaan; MILLET, Jose. Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing. **Sustainability**, v. 12, n. 19, p. 7848, set. 2020. https://doi.org/10.3390/su12197848

GRIOL-BARRES, Israel; MILLA, Sergio; CEBRIAN, Antonio ; MANSOORI, Yashar; MILLET, Jose. Variational Quantum Circuits for Machine Learning. An Application for the Detection of Weak Signals. **Applied Sciences**, v. 11, n. 14, p. 6427, jul. 2021. https://doi.org/10.3390/app11146427

GROVER, R. B. The Relationship between Science and Technology and Evolution in Methods of Knowledge Production. **Indian Journal of History of Science**, v. 54, n. 1, mar. 2019. https://doi.org/10.16943/ijhs/2019/v54i1/49597

GUO, Ying; XU, Chen; HUANG, Lu; PORTER, Alan L. Empirically Informing a Technology Delivery System Model for an Emerging Technology: Illustrated for Dye-Sensitized Solar Cells: Technology Delivery System for Dye-Sensitized Solar Cells. **R&D Management**, v. 42, n. 2, p. 133–49, mar. 2012. https://doi.org/10.1111/j.1467-9310.2012.00674.x

GUPTA, V.; LEHAL, G. S. A Survey of Text Mining Techniques and Applications. **Journal of Emerging Technologies in Web Intelligence**, v. 1, n. 1, p. 60-76, 2009.

HAEGEMAN, Karel; MARINELLO, Elisabetta; SCAPOLO, Fabiana; RICCI, Andrea; SOKOLOV, Alexander. Quantitative and Qualitative Approaches in Future-Oriented Technology Analysis (FTA): From Combination to Integration?. **Technological Forecasting and Social Change**, v. 80, n. 3, p. 386–97, mar. 2013. https://doi.org/10.1016/j.techfore.2012.10.002

HALLIDAY, Michael Alexander K.. **The Language of Science**. Collected Works of MAK Halliday, Continuum, London, v.5, 2004.

HARRIS, S. Dyer; ZEISLER, Steven. Weak signals: Detecting the next big thing. **The Futurist**, v. 36, n. 6, p. 21, nov/dez. 2002.

HARRIS, Z. S.. Distributional structure. **Word**, [S.L.], v.10, n. 2-3, p.146-162, 1954. http://dx.doi.org/10.1080/00437956.1954.11659520

HEINONEN, Sirkka; KARJALAINEN; RUOTSALAINEN, Juho; STEINMÜLLER, Karlheinz. Surprise as the New Normal – Implications for Energy Security. **European Journal of Futures Research**, v. 5, n. 1, p. 12, dez. 2017. https://doi.org/10.1007/s40309-017-0117-5

HEVNER A.; CHATTERJEE, S.Design Science Research in Information Systems. **Design Research in Information Systems**. Integrated Series in Information Systems, v. 22, 2010. https://doi.org/10.1007/978-1-4419-5653-8_2

HEVNER, A.R.; MARCH, S.T; PARK, J; RAM, S. Design Science in Information Systems Research. **MIS Quarterly**, v. 28, n. 1, p. 75, 2004. http://dx.doi.org/10.2307/25148625

HILTUNEN, E.. Was it a wild card or just our blindness to gradual change?. **Journal of Futures Studies**, [S.L.], v.11, n.2, p.61-74, 2006.

HILTUNEN, Elina. The Future Sign and Its Three Dimensions. **Futures**, v. 40, n. 3, p. 247–60, abr. 2008. https://doi.org/10.1016/j.futures.2007.08.021

HIRST, R. Scientific Jargon, Good and Bad. **Journal of Technical Writing and Communication**, [S.L.], v. 33, n. 3, p. 201–229, 2003. https://doi.org/10.2190/J8JJ-4YD0-4R00-G5N0

HIRSCHBERG, Julia; MANNING, Christopher D.. Advances in Natural Language Processing. **Science**, v. 349, n. 6245, p. 261–66, jul.2015. https://doi.org/10.1126/science.aaa8685

HOFFER, Bates L. Language borrowing and the indices of adaptability and receptivity. **Intercultural communication studies**, v. 14, n. 2, p.53, 2005.

HOLOPAINEN, Mari; TOIVONEN, Marja. Weak signals: ansoff today. **Futures**, [S.L.], v. 44, n. 3, p. 198-205, abr. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2011.10.002

HOMBAIAH, Spurthi Amba; CHEN, Tao; ZHANG, Mingyang; BENDERSKY, Michael; NAJORK, Marc. Dynamic Language Models for Continuously Evolving Content. **In Proceedings Of The 27Th Acm Sigkdd Conference On Knowledge Discovery & Data Mining**, [S.L.], p. 2514-2524, 14 ago. 2021. ACM. http://dx.doi.org/10.1145/3447548.3467162

HOWARD J.; RUDER, S. Universal Language Model Fine-tuning for Text Classification. **In proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, Melbourne-Australia, p. 328–339, jul. 2018. Association for Computational Linguistics. arXiv:1801.06146

HU, Ziniu; CHEN, Ting; CHANG, Kai-Wei; SUN, Yizhou. Few-Shot Representation Learning for Out-Of-Vocabulary Words. **In Proceedings Of The 57Th Annual Meeting Of The Association For Computational Linguistics**, [S.L.], p. 4102-4112, 2019. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/p19-1402

HUANG, Lu; ZHANG, Yi; GUO, Ying; ZHU, Donghua; PORTER, Alan L. Four Dimensional Science and Technology Planning: A New Approach Based on Bibliometrics and Technology Roadmapping. **Technological Forecasting and Social Change**, v. 81, p. 39–48, jan. 2014. https://doi.org/10.1016/j.techfore.2012.09.010

HUANG, Ying; SCHUEHLE, Jannik; PORTER, Alan L.; YOUTIE, Jan. A Systematic Method to Create Search Strategies for Emerging Technologies Based on the Web of Science: Illustrated for 'Big Data.' **Scientometrics**, v. 105, n. 3, p. 2005–22, dez. 2015. https://doi.org/10.1007/s11192-015-1638-y

HUSSAIN, M.; TAPINOS, E.; KNIGHT, L.. Scenario-driven roadmapping for technology foresight. **Technological Forecasting And Social Change**, [S.L.], v. 124, p. 160-177, nov. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2017.05.005

HWANG, Seonho; SHIN, Juneseuk. Extending Technological Trajectories to Latest Technological Changes by Overcoming Time Lags. **Technological Forecasting and Social Change**, v. 143, p. 142–53, jun. 2019. https://doi.org/10.1016/j.techfore.2019.04.013

I'ANSON, Ian. Scientometric analysis of the emerging technology landscape. **Qualitative and Quantitative Methods in Libraries**, [S.l.], v. 5, n. 1, p. 1-10, mai. 2017. Disponível em: http://www.qqml-journal.net/index.php/qqml/article/view/23

IDEN, Jon; METHLIE, Leif B.; CHRISTENSEN, Gunnar E . The Nature of Strategic Foresight Research: A Systematic Literature Review. **Technological Forecasting and Social Change**, v. 116, p. 87–97, mar. 2017. https://doi.org/10.1016/j.techfore.2016.11.002

ILMOLA, Leena; KUUSI, Osmo. Filters of Weak Signals Hinder Foresight: Monitoring Weak Signals Efficiently in Corporate Decision-Making. **Futures**, v. 38, n. 8, p. 908–24, out. 2006. https://doi.org/10.1016/j.futures.2005.12.019

ISO 704:2009, Terminology work – Principles and methods, International Organization for Standardization (ISO), 3rd ed., 2009. Disponível em: https://www.iso.org/standard/38109.html

JANISSEK-MUNIZ, Raquel; LESCA, Humbert; FREITAS, Henrique. Inteligência Estratégica Antecipativa e Coletiva para Tomada de Decisão. **Revista Inteligência Competitiva**, v.1, n.1, p.102-127, 2011.

JAMET, Denis; TERRY, Adeline. Introduction. **Lexis**, [*s. l*], v. 12, p. 1-4, dez. 2018. https://doi.org/10.4000/lexis.2521

JAWAHAR, Ganesh; SAGOT, Benoît; SEDDAH, Djamé. What Does BERT Learn about the Structure of Language? **In Proceedings Of The 57Th Annual Meeting Of The Association For Computational Linguistics**, [S.L.], p. 3651-3657, 2019. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/p19-1356

JOHNSTON, Ron; CAGNIN, Cristiano. The influence of future-oriented technology analysis: addressing the cassandra challenge. **Futures**, [S.L.], v. 43, n. 3, p. 313-316, abr. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2010.11.008

JOUNG, Junegak; KIM, Kwangsoo. Monitoring Emerging Technologies for Technology Planning Using Technical Keyword Based Analysis from Patent Data. **Technological Forecasting and Social Change**, v. 114, p. 281–92, jna, 2017. https://doi.org/10.1016/j.techfore.2016.08.020

KAJIKAWA, Yuya, YOSHIKAWA, Junta; TAKEDA, Yoshiyuki; MATSUSHIMA, Katsumori. Tracking Emerging Technologies in Energy Research: Toward a Roadmap for Sustainable Energy. **Technological Forecasting and Social Change**, v. 75, n. 6, p. 771–82, jul. 2008. https://doi.org/10.1016/j.techfore.2007.05.005

KATSURAI, Marie; ONO, Shunsuke. TrendNets: mapping emerging research trends from dynamic co-word networks via sparse representation. **Scientometrics**, [S.L.], v. 121, n. 3, p. 1583-1598, 18 out. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11192-019-03241-6

KAYSER, Victoria; BLIND, Knut. Extending the Knowledge Base of Foresight: The Contribution of Text Mining. **Technological Forecasting and Social Change**, v. 116, p. 208–15, mar. 2017. https://doi.org/10.1016/j.techfore.2016.10.017

KELLER, Jonas; GRACHT, Heiko A. von der. The Influence of Information and Communication Technology (ICT) on Future Foresight Processes — Results from a Delphi

Survey. **Technological Forecasting and Social Change**, v. 85, p. 81–92, jun. 2014. https://doi.org/10.1016/j.techfore.2013.07.010

KEYES, Ralph. **The Hidden History of Coined Words**. New York: Oxford University Press, 2021.

KHODAK, Mikhail; SAUNSHI, Nikunj; LIANG, Yingyu; MA, Tengyu; STEWART, Brandon; ARORA, Sanjeev. A La Carte *Embedding*: cheap but effective induction of semantic feature vectors. **In Proceedings Of The 56Th Annual Meeting Of The Association For Computational Linguistics (Volume 1**: Long Papers), [S.L.], p. 12-22, 2018. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/p18-1002

KHUN, Thomas. **The Structure of Scientific Revolutions**, University Chicago Press, Chicago, IL, USA, 1996.

KIM, Gabjo; BAE, Jinwoo. A Novel Approach to Forecast Promising Technology through Patent Analysis. **Technological Forecasting and Social Change**, v. 117, p. 228–37, abr. 2017. https://doi.org/10.1016/j.techfore.2016.11.023

KIM, Mujin; PARK, Youngjin; YOON, Janghyeok. Generating Patent Development Maps for Technology Monitoring Using Semantic Patent-Topic Analysis. **Computers & Industrial Engineering**, v. 98, p. 289–99, ago. 2016. https://doi.org/10.1016/j.cie.2016.06.006

KIM, Seonho; KIM, You-Eil; BAE, Kuk-Jin; CHOI, Sung-Bae; PARK, Jong-Kyu; KOO, Young-Duk; PARK, Young-Wook; CHOI, Hyun-Kyoo; KANG, Hyun-Moo; HONG, Sung-Wha. NEST: A Quantitative Model for Detecting Emerging Trends Using a Global Monitoring Expert Network and Bayesian Network. **Futures**, v. 52, p. 59–73, aug. 2013. https://doi.org/10.1016/j.futures.2013.08.004

KIM, Jieun; LEE, Changyong. Novelty-focused weak signal detection in futuristic data: assessing the rarity and paradigm unrelatedness of signals. **Technological Forecasting And Social Change**, [S.L.], v. 120, p. 59-76, jul. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2017.04.006

KIM, Jieun; PARK, Yongtae; LEE, Youngjo. A Visual Scanning of Potential Disruptive Signals for Technology Roadmapping: Investigating Keyword Cluster, Intensity, and Relationship in Futuristic Data. **Technology Analysis & Strategic Management**, v. 28, n. 10, p. 1225–46, nov. 2016. https://doi.org/10.1080/09537325.2016.1193593

KIM, Tae San; SOHN, So Young. Machine-Learning-Based Deep Semantic Analysis Approach for Forecasting New Technology Convergence. **Technological Forecasting and Social Change**, v. 157, p. 120095, ago. 2020. https://doi.org/10.1016/j.techfore.2020.120095

KLEINBERG, Jon. Bursty and hierarchical structure in streams. **Data mining and knowledge discovery**, v.7, n. 4, p. 373-397, 2003. https://doi.org/10.1023/A:1024940629314

KONTOSTATHIS, April; GALITSKY, Leon M.; POTTENGER, William M.; ROY, Soma; PHELPS, Daniel J.. A survey of emerging trend detection in textual data mining. **In Survey of text mining**, p. 185-224. Springer, New York, NY, 2004.

KOSHLAKOV, D.; KHOKHLOVA, M.; TSAREVA, G.; GARBUZOVA, G. Eponyms in science terms (Epistemological aspect). **In Proceedings of the 2019 International Scientific**

Conference: "Achievements and Perspectives of Philosophical Studies" (APPSCONF), [S.L], nov. 2019. http://doi.org/10.1051/shsconf/20197201016

KRIEGER, Maria da Graça. Terminologia técnico-científica: políticas lingüísticas e Mercosul. **Ciência e Cultura**, v. 58, n. 2, p. 45-48, 2006.

KUMAR, Kamalesh; SUBRAMANIAN, Ram; STRANDHOLM, Karen. Competitive Strategy, Environmental Scanning and Performance: A Context Specific Analysis of Their Relationship. **International Journal of Commerce and Management**, v. 11, n. 1, p. 1–33, jan. 2001. https://doi.org/10.1108/eb047413

KUOSA, Tuomo. Futures Signals Sense-Making Framework (FSSF): A Start-up Tool to Analyse and Categorise Weak Signals, Wild Cards, Drivers, Trends and Other Types of Information. **Futures**, v. 42, n. 1, p. 42–48, fev. 2010. https://doi.org/10.1016/j.futures.2009.08.003

KURFESS, Thomas R. **Robotics and Automation Handbook**. Taylor & Francis. ISBN 9780849318047, 2005.

KUTUZOV, Andrey; ØVRELID, Lilja; SZYMANSKI, Terrence; VELLDAL, Erik. Diachronic Word Embeddings and Semantic Shifts: A Survey. **In Proceedings of the 27th International Conference on Computational Linguistics**, p. 1384–1397, 2018. https://doi.org/10.48550/ARXIV.1806.03537.

KUUSI, Osmo; MEYER, Martin. Technological Generalizations and Leitbilder—the Anticipation of Technological Opportunities. **Technological Forecasting and Social Change**, v. 69, n. 6, p. 625–39, jul. 2002. https://doi.org/10.1016/S0040-1625(02)00182-8

KWON, Lee-Nam; PARK, Jun-Hwan; MOON, Yeong-Ho; LEE, Bangrae; SHIN, Youngho; KIM, Young-Kuk. Weak signal detecting of industry convergence using information of products and services of global listed companies - focusing on growth engine industry in South Korea -. **Journal Of Open Innovation**: Technology, Market, and Complexity, [S.L.], v. 4, n. 1, p. 1-19, 27 mar. 2018. MDPI AG. http://dx.doi.org/10.1186/s40852-018-0083-6

LACERDA, D. P.; DRESCH A.; PROENÇA, A.; ANTUNES Jr., J. A. V. Design Science Research: Método de Pesquisa Para a Engenharia de Produção. **Gestão & Produção**, v. 20, n. 4, p. 741–61, nov. 2013. https://doi.org/10.1590/S0104-530X2013005000014

LANSING, J. Stephen. Complex Adaptive Systems. Annual Review Of Anthropology, [S.L.], v. 32, n. 1, p. 183-204, out. 2003. Annual Reviews. http://dx.doi.org/10.1146/annurev.anthro.32.061002.093440

LAHOTI, Geet; PORTER, Alan L.; ZHANG, Chuck; YOUTIE, Jan; WANG, Ben. Tech Mining to Validate and Refine a Technology Roadmap. **World Patent Information**, v. 55, p. 1–18, dez. 2018. https://doi.org/10.1016/j.wpi.2018.07.003

LAURIOLA, Ivano, LAVELLI, Alberto; AIOLLI, Fabio. An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools. **Neurocomputing**, v. 470, p. 443–56, jan. 2022. https://doi.org/10.1016/j.neucom.2021.05.103

LEE, Changyong. A Review of Data Analytics in Technological Forecasting. **Technological Forecasting and Social Change**, v. 166, p. 120646, mai. 2021. https://doi.org/10.1016/j.techfore.2021.120646

LEE, Changyong; KWON, Ohjin; KIM, Myeongjung; KWO, Daeil. Early Identification of Emerging Technologies: A Machine Learning Approach Using Multiple Patent Indicators. **Technological Forecasting and Social Change**, v. 127, p. 291–303, fev. 2018. https://doi.org/10.1016/j.techfore.2017.10.002

LEE, D.; LEE, H. Mapping the Characteristics of Design Research in Social Sciences. **Archives of Design Research**, v.32, n.4, p.39-51, 2019. http://dx.doi.org/10.15187/adr.2019.11.32.4.39

LEE, June Young; AHN, Sejung; KIM, Dohyun. Deep Learning-Based Prediction of Future Growth Potential of Technologies. PLOS ONE, v. 16, n. 6, p. E0252753, jun. 2021. https://doi.org/10.1371/journal.pone.0252753

LEE, Young-Joo; PARK, Ji-Young. Identification of Future Signal Based on the Quantitative and Qualitative Text Mining: A Case Study on Ethical Issues in Artificial Intelligence. **Quality & Quantity**, v. 52, n. 2, p. 653–67, mar. 2018. https://doi.org/10.1007/s11135-017-0582-8

LEE, J.; YOON, W.; KIM, S.; KIM, D.; KIM, S.; Ho So, C.; KANG, J.. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, [S.L.], v. 36, n. 4, p. 1234-1240, fev 2020. https://doi.org/10.1093/bioinformatics/btz682

LESCA, Humbert. The crucial problem of the strategic probe: the construction of the ‘ puzzle’. **Report - Centre d'Etudes et de Recherches Appliquees a la Gestion (CERAG)**, France, 1995.

LESCA, Humbert; LESCA, Nicolas. **Weak Signals for Strategic Intelligence: Anticipation Tool for Managers**. John Wiley & Sons, Inc, 2011. https://doi.org/10.1002/9781118602775

LI, Munan. An Exploration to Visualise the Emerging Trends of Technology Foresight Based on an Improved Technique of Co-Word Analysis and Relevant Literature Data of WOS. **Technology Analysis & Strategic Management**, v. 29, n. 6, p. 655–71, jul. 2017. https://doi.org/10.1080/09537325.2016.1220518

LI, Munan. Capturing the Risk Signals for a Specific Emerging Technology: An Integrated Framework of Text Mining. **IEEE Transactions on Engineering Management**, v. 68, n. 5, p. 1245–58, out. 2021. https://doi.org/10.1109/TEM.2019.2930335

LI, Munan; CHU, Yanqun. Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. **Journal Of Information Science**, [S.L.], v. 43, n. 6, p. 725-741, 1 set. 2016. SAGE Publications. http://dx.doi.org/10.1177/0165551516661914

LI, Munan; PORTER, Alan L.; SUOMINEN, Arho. Insights into relationships between disruptive technology/innovation and emerging technology: a bibliometric perspective. **Technological Forecasting And Social Change**, [S.L.], v. 129, p. 285-296, abr. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2017.09.032

LI, Qing; ZHANG, Huaige; HONG, Xianpei. Knowledge Structure of Technology Licensing Based on Co-Keywords Network: A Review and Future Directions. **International Review of Economics & Finance**, v. 66, p. 154–65, mar. 2020. https://doi.org/10.1016/j.iref.2019.11.007

LI, Xin; FAN, Mingjie; ZHOU, Yuan; FU, Jing; YUAN, Fei; HUANG, Lucheng. Monitoring and Forecasting the Development Trends of Nanogenerator Technology Using Citation Analysis and Text Mining. **Nano Energy**, v. 71, p. 104636, mai. 2020. https://doi.org/10.1016/j.nanoen.2020.104636

LI, Xin; XIE, Qianqian; DAIM, Tugrul; HUANG, Lucheng. Forecasting Technology Trends Using Text Mining of the Gaps between Science and Technology: The Case of Perovskite Solar Cell Technology. **Technological Forecasting and Social Change**, v. 146, p. 432–49, set. 2019. https://doi.org/10.1016/j.techfore.2019.01.012

LI, Xiaotao; YOU, Shujuan; NIU, Yawen; CHEN, Wai. Learning *Embeddings* for Rare Words Leveraging Internet Search Engine and Spatial Location Relationships. **In Proceedings Of *Sem 2021**: The Tenth Joint Conference on Lexical and Computational Semantics, [S.L.], p. 278-287, 2021. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2021.starsem-1.26

LI, Yuan; DAI, Jie; FAN, Xiao-Lin; ZHAO, Yu-Hai; WANG, Guo-Ren. I/O Efficient Early Bursting Cohesive Subgraph Discovery in Massive Temporal Networks. **Journal of Computer Science and Technology**, v. 37, n. 6, p. 1337–55, dez. 2022. https://doi.org/10.1007/s11390-022-2367-3.

LI, Yuan; LIU, Jinsheng; ZHAO, Huiqun; SUN, Jing; ZHAO, Yuhai; WANG, Guoren. Efficient Continual Cohesive Subgraph Search in Large Temporal Graphs. **World Wide Web**, v. 24, n. 5, p. 1483–509, set. 2021. https://doi.org/10.1007/s11280-021-00917-z.

LIANG, Hong; SUN, Xiao; SUN, Yunlei; GAO, Yuan. Text feature extraction based on deep learning: a review. **Eurasip Journal On Wireless Communications And Networking**, [S.L.], v. 2017, n. 1, p. 1-12, dez. 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s13638-017-0993-1

LIAO, Shu-Hsien; CHU, Pei-Hui; HSIAO, Pei-Yuan. Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. **Expert Systems with Applications**, v. 39, n. 12, p. 11303–11, set. 2012. https://doi.org/10.1016/j.eswa.2012.02.063

LINSTONE, Harold A. On Terminology. **Technological Forecasting and Social Change**, v. 77, n. 9, p. 1426–27, nov. 2010. https://doi.org/10.1016/j.techfore.2010.08.002

LINSTONE, Harold A. Three Eras of Technology Foresight. **Technovation**, v. 31, n. 2–3, p. 69–76, fev. 2011. https://doi.org/10.1016/j.technovation.2010.10.001

LINSTONE, Harold A.; TUROFF, Murray . Delphi: A Brief Look Backward and Forward. **Technological Forecasting and Social Change**, v. 78, n. 9, p. 1712–19, nov. 2011. https://doi.org/10.1016/j.techfore.2010.09.011

LIU, Huailan; CHEN, Zhiwang; TANG, Jie; ZHOU, Yuan; LIU, Sheng. Mapping the Technology Evolution Path: A Novel Model for Dynamic Topic Detection and Tracking. **Scientometrics**, v. 125, n. 3, p. 2043–90, dec. 2020. https://doi.org/10.1007/s11192-020-03700-5

LIU, Jiaying; TANG, Tao; WANG, W.; XU, Bo; KONG, Xiangjie; XIA, Feng. A Survey of Scholarly Data Visualization. **IEEE Access**, v. 6, p. 19205–21, 2018. https://doi.org/10.1109/ACCESS.2018.2815030

LIU, Xuanming; GE, Tingjian; WU, Yinghui. A Stochastic Approach to Finding Densest Temporal Subgraphs in Dynamic Graphs. **IEEE Transactions on Knowledge and Data Engineering**, p. 1–1, 2020. https://doi.org/10.1109/TKDE.2020.3025463.

LUFT, Joe; INGHAM, Harry. The Johari window, a graphic model of interpersonal awareness, **Proceedings of the Western Training Laboratory in Group Development**. Los Angeles: UCLA, 1955.

LUNDVALL Bengt-Åke; BORRÁS, Susana. Science, technology and innovation policy. **The Oxford handbook of innovation**, p. 599-631, 2005.

MA, Shuai; HU, Renjun; WANG, Luoshu; LIN, Xuelian; HUAI, Jinpeng. An Efficient Approach to Finding Dense Temporal Subgraphs. **IEEE Transactions on Knowledge and Data Engineering**, v. 32, n. 4, p. 645–58, abr. 2020. https://doi.org/10.1109/TKDE.2019.2891604.

MA, Tingting; PORTER, Alan L.; GUO, Ying; READY, Jud; XU, Chen; GAO, Lidan. A Technology Opportunities Analysis Model: Applied to Dye-Sensitised Solar Cells for China. **Technology Analysis & Strategic Management**, v. 26, n. 1, p. 87–104, jan. 2014. https://doi.org/10.1080/09537325.2013.850155

MACHADO, Guilherme Bertoni; KRAEMER, Rodrigo; DANDOLINI, Gertrudes Aparecida; SOUZA, João Artur de; TODESCO, José Leomar. PERSPECTIVAS DE PESQUISA SOBRE INTELIGÊNCIA ESTRATÉGICA ANTECIPATIVA E COLETIVA (IEAc) POR MEIO DA ANÁLISE DE SENTIMENTO: UM CENÁRIO DIDÁTICO DE USO. **Perspectivas em Gestão & Conhecimento**, v. 10, n. 1, mai. 2020. https://doi.org/10.21714/2236-417X2020v10n1p152

MACIEL, Daniel; ARTESE, Leticia S.; GONÇALVES, Alexandre L.. WORD EMBEDDING FOR UNKNOWN WORDS: ADDING NEW WORDS INTO BERT'S VOCABULARY. **19th CONTECSI International Conference on Information Systems and Technology Management, TECSI**, 2022. https://doi.org/10.5748/19CONTECSI/PSE/DSC/7035

MAGRUK, Andrzej. Analysis of Uncertainties and Levels of Foreknowledge in Relation to Major Features of Emerging Technologies—The Context of Foresight Research for the Fourth Industrial Revolution. **Sustainability**, v. 13, n. 17, p. 9890, set. 2021. https://doi.org/10.3390/su13179890

MAIER, H. R.; GUILLAUME, J.H.A.; DELDEN, H.van; RIDDELL, G.A.; HAASNOOT, M.; KAWAKKEL, J.H.. An Uncertain Future, Deep Uncertainty, Scenarios, Robustness and Adaptation: How Do They Fit Together? **Environmental Modelling & Software**, v. 81, p. 154–64, jul. 2016. https://doi.org/10.1016/j.envsoft.2016.03.014

MALONE, David; TURNER, James. The Merger of AOL and Time Warner: a case study. Journal of the International Academy for case studies, v. 16, n. 7, p. 103-109, jan. 2010.

MARCH, S. T.; SMITH, G. F. Design and natural science research in Information Technology. **Decision Suport Systems**, v. 15, p. 251-266, 1995. http://dx.doi. org/10.1016/0167-9236(94)00041-2

MARTIN, Ben R.. Foresight in science and technology. **Technology Analysis & Strategic Management**, [S.L.], v. 7, n. 2, p. 139-168, jan. 1995. Informa UK Limited. http://dx.doi.org/10.1080/09537329508524202

MARTIN, Ben R. The Origins of the Concept of 'Foresight' in Science and Technology: An Insider's Perspective. **Technological Forecasting and Social Change**, v. 77, n. 9, p. 1438–47, nov. 2010. https://doi.org/10.1016/j.techfore.2010.06.009

MCCARTHY, Ian P.. Technology management a complex adaptive systems approach. **International Journal Of Technology Management**, [S.L.], v. 25, n. 8, p. 728, 2003. Inderscience Publishers. http://dx.doi.org/10.1504/ijtm.2003.003134

MEISSNER, Dirk; GOKHBERG, Leonid; SOKOLOV, Alexander Sokolov. **Science, technology and innovation policy for the future: potentials and limits of foresight studies**. New York, Dordrecht, London, Heidelberg: Springer, 2013.

MENDONÇA, Sandro; CARDOSO, Gustavo; CARAÇA, João. The strategic strength of weak signal analysis. **Futures**, [S.L.], v. 44, n. 3, p. 218-228, abr. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2011.10.004

MENEGHETTI, F.K. Pragmatismo e os pragmáticos nos estudos organizacionais. Cad. **EBAPE.BR**, v. 5, n. 1, p. 1-13, 2007. https://doi.org/10.1590/S1679-39512007000100005

MIAO, Hong; GUO, Xin; YUAN, Fei. Research on Identification of Potential Directions of Artificial Intelligence Industry From the Perspective of Weak Signal. **IEEE Transactions on Engineering Management**, p. 1–16, 2021. https://doi.org/10.1109/TEM.2021.3123639

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.S.; DEAN, J. Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, [S.L.], p. 3111-3119, 2013.

MILES, Ian. The Development of Technology Foresight: A Review. Technological Forecasting and Social Change, v. 77, n. 9, p. 1448–56, nov. 2010. https://doi.org/10.1016/j.techfore.2010.07.016

MOEHRLE, Martin G.; CAFEROGLU, Hüseyin. Technological Speciation as a Source for Emerging Technologies. Using Semantic Patent Analysis for the Case of Camera Technology. **Technological Forecasting and Social Change**, v. 146, p. 776–84, set. 2019. https://doi.org/10.1016/j.techfore.2018.07.049

MOHAMMADI, Mohsen; EIVAZI, Mohammad Rahim; SAJJADI, Jafar. Wildcards – natural and artificial: the combination of a panel of experts and fuzzy topsis. **Foresight**, [S.L.], v. 19, n. 1, p. 15-30, 13 mar. 2017. Emerald. http://dx.doi.org/10.1108/fs-08-2016-0040

MOHAMMED, Athar Hussein; ALI H. Ali. Survey of BERT (Bidirectional Encoder Representation Transformer) Types. **Journal of Physics: Conference Series**, v. 1963, n. 1, p. 012173, jul.2021. https://doi.org/10.1088/1742-6596/1963/1/012173

MOLONTAY, Roland; NAGY, Marcell. Two Decades of Network Science: As Seen through the Co-Authorship Network of Network Scientists. **Proceedings International Conference on Advances in Social Networks Analysis and Mining** (IEEE/ACM), p. 578–83, 2019. https://doi.org/10.1145/3341161.3343685

MOMENI, Abdolreza; ROST, Katja. Identification and Monitoring of Possible Disruptive Technologies by Patent-Development Paths and Topic Modeling. **Technological Forecasting and Social Change**, v. 104, p. 16–29, mar. 2016 https://doi.org/10.1016/j.techfore.2015.12.003

MOON, S.; OKAZAKI, N..PatchBERT: Just-in-Time, Out-of-Vocabulary Patching. **In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, [S.L.], p.7846–7852, 2020. http://dx.doi.org/10.18653/v1/2020.emnlp-main.631

MOONEY, R. J.; NAHM, U. Y. Text Mining with Information Extraction. **Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium**. Bloemfontein, South Africa: Van Schaik Pub.: 141-160 p. 2005.

MORESI, Eduardo Amadeu Dutra. Proposta de abordagem para a detecção de sinais fracos. **Parcerias Estratégicas**, v. 23, n. 46, p. 09-28, 2019.

MORIANO, Pablo; FINKE, Jorge; AHN, Yong-Yeol. Community-Based Event Detection in Temporal Networks. **Scientific Reports**, v. 9, n. 1, p. 4358, mar. 2019. https://doi.org/10.1038/s41598-019-40137-0.

MORO, Alberto; BOELMAN, Elisa; JOANNY, Geraldine; GARCIA, Juan Lopez. A Bibliometric-Based Technique to Identify Emerging Photovoltaic Technologies in a Comparative Assessment with Expert Review. **Renewable Energy**, v. 123, p. 407–16, ago. 2018. https://doi.org/10.1016/j.renene.2018.02.016

MÜHLROTH, Christian; GROTTKE, Michael. A systematic literature review of mining weak signals and trends for corporate foresight. **Journal Of Business Economics**, [S.L.], v. 88, n. 5, p. 643-687, 19 mar. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11573-018-0898-4

NADKARNI, Prakash M.; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. Natural Language Processing: An Introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–51, set. 2011. https://doi.org/10.1136/amiajnl-2011-000464

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, v. 45, n. 2, p. 167–256, jan. 2003. https://doi.org/10.1137/S003614450342480

NEWMAN, M. E. J. Modularity and Community Structure in Networks. **Proceedings of the National Academy of Sciences**, v. 103, n. 23, p. 8577–82, jun. 2006. https://doi.org/10.1073/pnas.0601602103

NEWMAN, M. E. J.; GIRVAN, M. Finding and Evaluating Community Structure in Networks. Physical Review E, v. 69, n. 2, p. 026113, fev. 2004. https://doi.org/10.1103/PhysRevE.69.026113

NEWMAN, Nils C.; PORTER, Alan L.; NEWMAN, David; TRUMBACH, Cherie Courseault; BOLAN, Stephanie D. Comparing Methods to Extract Technical Content for Technological Intelligence. **Journal of Engineering and Technology Management**, v. 32, p. 97–109, abr. 2014, https://doi.org/10.1016/j.jengtecman.2013.09.001

NIINILUOTO, Ilkka. Futures studies: science or art?. **Futures**, [S.L.], v. 33, n. 5, p. 371-377, jun. 2001. Elsevier BV. http://dx.doi.org/10.1016/s0016-3287(00)00080-x

NONAKA, Ikujiro. The Knowledge-Creating Company. **Harvard Business Review**, [*s. l*], v. 69, n. 6, p. 96-104, nov. 1991.

OGAWA, Takaya; KAJIKAWA, Yuya. Assessing the Industrial Opportunity of Academic Research with Patent Relatedness: A Case Study on Polymer Electrolyte Fuel Cells. **Technological Forecasting and Social Change**, v. 90, p. 469–75, jan. 2015. https://doi.org/10.1016/j.techfore.2014.04.002

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **Manual de Oslo: proposta de diretrizes para coleta e interpretação de dados sobre inovação tecnológica**. São Paulo, SP: FINEP, 1997.

PARK, Chankook; CHO, Seunghyun. Future Sign Detection in Smart Grids through Text Mining. **Energy Procedia**, v. 128, p. 79–85, set. 2017. https://doi.org/10.1016/j.egypro.2017.09.018

PARK, Inchae; YOON, Byungun. Technological Opportunity Discovery for Technological Convergence Based on the Prediction of Technology Knowledge Flow in a Citation Network. **Journal of Informetrics**, v. 12, n. 4, p. 1199–222, nov. 2018. https://doi.org/10.1016/j.joi.2018.09.007

PATENT INFORMATICS TEAM. An Analysis of Worldwide Patent Filings Relating to Graphene, **Intellectual Property Office**, UK, 2011. Disponível em: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/312521/informatic-graphene.pdf

PEIRCE, C.S. **The Collected Papers of Charles Sanders Peirce**. Harvard University Press, Cambridge, v. 1-8, 1931-1958.

PEIRCE, Charles Sanders. **Semiótica e filosofia**. São Paulo : Cultrix, 1972.

PEFFERS, K.; TUUNANEN, T; ROTHENBERGER, M. A.; CHATTERJEE, S. A Design Science Research Methodology for Information Systems Research, **Journal of Management Information Systems**, v. 24, n.3, p. 45-77, 2007. http://dx.doi.org/10.2753/MIS0742-1222240302

PENNINGTON, J.; SOCHER, R.; MANNING, C.D.. Glove: Global Vectors for Word Representation. **In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, [S.L.], 2014. http://dx.doi.org/10.3115/v1/D14-1162

PETER, Marc K.; JARRATT, Denise G.. The practice of foresight in long-term planning. **Technological Forecasting And Social Change**, [S.L.], v. 101, p. 49-61, dez. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2013.12.004

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L.. Deep contextualized word representations. **In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)**, [S.L.], 2018. http://dx.doi.org/10.18653/v1/N18-1202

PETERSEN, J. L.. **Out of the blue: How to anticipate big future surprises**. Lanham, Md: Madison Books, 1999.

PIETREWICZ, Lesław. Technology, Business Models and Competitive Advantage in the Age of Industry 4.0. **Problemy Zarzadzania**, v. 2, n. 82, p. 32–52, mai. 2019. https://doi.org/10.7172/1644-9584.82.2

PIETROBELLI, Carlo; PUPPATO, Fernanda. Technology Foresight and Industrial Strategy. **Technological Forecasting and Social Change**, v. 110, p. 117–25, set. 2016. https://doi.org/10.1016/j.techfore.2015.10.021

PONOMAREV, Ilya V., WILLIAMS, Duane E.; HACKETT, Charles, J.; SCHNELL, Joshua D.; HAAK, Laurel L. Predicting Highly Cited Papers: A Method for Early Detection of Candidate Breakthroughs. **Technological Forecasting and Social Change**, v. 81, p. 49–55, jan, 2014. https://doi.org/10.1016/j.techfore.2012.09.017

PONTES, A. L. Terminologia Científica: O que é e como se faz. **Revista de Letras**, [S.L.], v. 19, n. 1, p. 11, 1997.

PORTER, Alan L.; ROPER, A. Thomas; MASON, Thomas W.; ROSSINI, Frederick A.; BANKS, Jerry; WIEDERHOLT, Bradley J. Forecasting and Management of Technology. John Wiley & Sons, Inc., 1991. https://doi.org/10.1002/9781118047989

PORTER, Alan.L., Future-oriented Technology Analyses: The Literature and Its Disciplines. **Knowing Tomorrow? How Science Deals with the Future**, Eburon Academic Publishers, Delft, 183-201, 2007.

PORTER, Alan L. Technology Foresight: Types and Methods. **International Journal of Foresight and Innovation Policy**, vl. 6, n. 1/2/3, p. 36, 2010. https://doi.org/10.1504/IJFIP.2010.032664

PORTER, Alan L.; CHIAVETTA, Denise; NEWMAN, Nils C. Measuring Tech Emergence: A Contest. **Technological Forecasting and Social Change**, v. 159, p. 120176, out. 2020. https://doi.org/10.1016/j.techfore.2020.120176

PORTER, Alan L., CUNNINGHAM, Scott. W. **Tech Mining: Exploiting New Technologies for Competitive Advantage**, vol. 29. John Wiley & Sons, 2004.

PORTER, Alan L., DETAMEPEL, Michael J.. Technology Opportunities Analysis. **Technological Forecasting and Social Change**, v. 49, n. 3, p. 237–55, jul. 1995. https://doi.org/10.1016/0040-1625(95)00022-3

PORTER, Alan L., GARNER, Jon Gregory; CARLEY, Stephen; NEWMAN, Nils C. Emergence Scoring to Identify Frontier R&D Topics and Key Players. **Technological Forecasting and Social Change**, v. 146, p. 628–43, set. 2019. https://doi.org/10.1016/j.techfore.2018.04.016

PORTER, Alan L.; ROSSINI, Frederick A. Evaluation Designs for Technology Assessments and Forecasts. **Technological Forecasting and Social Change**, v. 10, n. 4, p. 369–80, jan. 1977. https://doi.org/10.1016/0040-1625(77)90033-6

PORTER, Alan.L.; ZHANG, Yi. Tech Mining of Science & Technology Information Resources for Future-Oriented Technology Analyses. **Futures Research Methodology** Version 3.1, The Millennium Project, Washington, DC, 2015.

POURGHASEM, J.; KARIMI, S.; EDALATPANAH, S.. A Survey of Voice Over Internet Protocol (VOIP) Technology. **International Journal Of Computer Mathematical Science And Applications (IJCMSA)**, v.6, n.3-4, p.53-62, 2012.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2ª ed. Novo Hamburgo: FEEVALE, 2013.

PURAO, S.; BALDWIN, C. Y.; HEVNER, A.; STOREY, V. C.; PRIES-HEJE, J.; SMITH, B.; ZHU, Y.. The Sciences of Design: Observations on an Emerging Field, **SSRN Electronic Journal**, 2008. https://doi.org/10.2139/ssrn.1281643

PURAO, Sandeep. Truth or Dare: The Ontology Question in Design Science Research, **Journal of Database Management**, v. 24, n. 3, p. 51-66, jul. 2013.

QIN, Hongchao; LI, Rong-Hua; WANG, Guoren; QIN, Lu; YUAN, Ye; ZHANG, Zhiwei. Mining bursting communities in temporal graphs. Nov. 2019. http://arxiv.org/abs/1911.02780

QIN, Hongchao; LI, Rong-Hua; YUAN, Ye; WANG, Guoren; QIN, Lu; ZHANG, Zhiwei. Mining Bursting Core in Large Temporal Graphs. **Proceedings of the VLDB Endowment**, v. 15, n. 13, p. 3911–23, set. 2022. https://doi.org/10.14778/3565838.3565845.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I.. Language models are unsupervised multitask learners. **OpenAI blog**, 1(8), 9, 2019.

RAAN, Anthony F. J. van. Sleeping Beauties in Science. **Scientometrics**, v. 59, n. 3, p. 467–72, 2004. https://doi.org/10.1023/B:SCIE.0000018543.82441.f1

RAAN, Anthony F. J. van. Sleeping Beauties Cited in Patents: Is There Also a Dormitory of Inventions? **Scientometrics**, v. 110, n. 3, p. 1123–56, mar. 2017. https://doi.org/10.1007/s11192-016-2215-8

RANAEI, S.; SUOMINEN, A.; PORTER, A.; CARLEY, S.. Evaluating technological emergence using text analytics: two case technologies and three approaches. **Scientometrics**, [S.L.], v.122, p. 215–247, 2020. https://doi.org/10.1007/s11192-019-03275-w

RANDVIIR, Edward P.; BROWSON; Dale A.C.; BANKS, Craig E.. A Decade of Graphene Research: Production, Applications and Outlook. **Materials Today**, v. 17, n. 9, p. 426–32, nov. 2014. https://doi.org/10.1016/j.mattod.2014.06.001

RESING, H. A.; MILLIKEN, J.; DOMINGUEZ, D. D.; ITON, L. E. The Orientation of Graphite Planes in Carbon Fibers via Carbon-l3 NMR Spectroscopy. Proceedings Biennial Conference on Carbon - Extended Abstracts and Program, p.84, 1985.

ROBERTSON, S.. Understanding inverse document frequency: on theoretical arguments for IDF. **Journal of Documentation**, [S.L.], v. 60, n. 5, p. 503-520, 2004. https://doi.org/10.1108/00220410410560582

ROHRBECK, René; BATTISTELLA, Cinzia; HUIZINGH, Eelko. Corporate foresight: an emerging field with a rich tradition. **Technological Forecasting And Social Change**, [S.L.], v. 101, p. 1-9, dez. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2015.11.002

ROHRBECK, René; SCHWARZ, Jan Oliver. The value contribution of strategic foresight: insights from an empirical study of large european companies. **Technological Forecasting And Social Change**, [S.L.], v. 80, n. 8, p. 1593-1606, out. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2013.01.004

ROSENBERG, Nathan. **Inside the Black Box: Technology and Economics**. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship, 1982.

ROSSEL, Pierre. Weak signals as a flexible framing space for enhanced management and decision-making. **Technology Analysis & Strategic Management**, [S.L.], v. 21, n. 3, p. 307-320, abr. 2009. Informa UK Limited. http://dx.doi.org/10.1080/09537320902750616

ROSSEL, Pierre. Beyond the obvious: examining ways of consolidating early detection schemes. **Technological Forecasting And Social Change**, [S.L.], v. 78, n. 3, p. 375-385, mar. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.techfore.2010.06.016

ROSSEL, Pierre. Early detection, warnings, weak signals and seeds of change: a turbulent domain of futures studies. **Futures**, [S.L.], v. 44, n. 3, p. 229-239, abr. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2011.10.005

ROTOLO, Daniele; HICKS, Diana; MARTIN, Ben R.. What Is an Emerging Technology? **Research Policy**, v. 44, n. 10, p. 1827–43, dez. 2015. https://doi.org/10.1016/j.respol.2015.06.006

ROUSSEAU, Pauline; CAMARA, Daniel; KOTZINOS, Dimitris. Weak Signal Detection and Identification in Large Data Sets: A Review of Methods and Applications, 2021 (preprint). https://doi.org/10.13140/RG.2.2.20808.24327

ROWE, Gene; WRIGHT, George. The Delphi Technique as a Forecasting Tool: Issues and Analysis. **International Journal of Forecasting**, v. 15, n. 4, p. 353–75, out. 1999. https://doi.org/10.1016/S0169-2070(99)00018-7

ROZENSHTEIN, Polina;BONCHI Francesco Bonchi; GIONIS, Aristides; SOZIO, Mauro; TATTI, Nikolaj. Finding Events in Temporal Networks: Segmentation Meets Densest Subgraph Discovery. **Knowledge and Information Systems**, v. 62, n. 4, p. 1611–39, abr. 2020. https://doi.org/10.1007/s10115-019-01403-9.

SACCOL, A. Z. Um retorno ao básico: Compreendendo os paradigmas de pesquisa e sua aplicação na pesquisa em administração. **Revista de Administração da Universidade Federal de Santa Maria**, v. 2, n. 2, p. 250-269, 2009.

SCHWARZ, Jan Oliver; KROEHL, Rixa; GRACHT, Heiko A. von der. Novels and Novelty in Trend Research — Using Novels to Perceive Weak Signals and Transfer Frames of Reference. **Technological Forecasting and Social Change**, v. 84, p. 66–73, mai. 2014. https://doi.org/10.1016/j.techfore.2013.09.007.

SALTON, G.; WONG, A.; YANG, C. S.. A vector space model for automatic indexing. **Communications Of The Acm**, [S.L.], v. 18, n. 11, p. 613-620, nov. 1975. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/361219.361220

SANTAELLA, Lucia. **Semiótica aplicada**. 2. ed. São Paulo: Cengage Learning, 2018. E-BOOK. ISBN 9788522126989.

SANTOS, Neri dos; RADOS, Gregório Jean Varvakis. **Fundamentos Teóricos de Gestão Do Conhecimento**. 1st ed., Florianópolis: Pandion, ISBN: 978-65-86527-01-8, 2020.

SARDAR, Ziauddin; SWEENEY, John A.. The Three Tomorrows of Postnormal Times. **Futures**, [S.L.], v. 75, p. 1-13, jan. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2015.10.004

SARICA, Serhad; LUO, Jianxi; WOOD, Kristin L.. TechNet: Technology Semantic Network Based on Patent Data. **Expert Systems with Applications**, v. 142, p. 112995, mar. 2020. https://doi.org/10.1016/j.eswa.2019.112995

SARITAS, Ozcan; BURMAOGLU, Serhat. The Evolution of the Use of Foresight Methods: A Scientometric Analysis of Global FTA Research Output. **Scientometrics**, v. 105, n. 1, p. 497–508, out. 2015. https://doi.org/10.1007/s11192-015-1671-x

SARITAS, Ozcan; SMITH, Jack E. The Big Picture – Trends, Drivers, Wild Cards, Discontinuities and Weak Signals. **Futures**, v. 43, n. 3, p. 292–312, abr. 2011. https://doi.org/10.1016/j.futures.2010.11.007

SAVOV, Pavel, JATOWT, Adam; NIELEK, Radoslaw. Identifying Breakthrough Scientific Papers. **Information Processing & Management**, v. 57, n. 2, p. 102168, mar. 2020. https://doi.org/10.1016/j.ipm.2019.102168

SHIBAIYAMA, Sotaro; YIN, Deyun; MATSUMOTO, Kuniko. Measuring Novelty in Science with Word *Embedding*. **PLOS ONE**, v. 16, n. 7, p. E0254034, jul. 2021. https://doi.org/10.1371/journal.pone.0254034

SCHICK, Timo; SCHÜTZE, Hinrich. Rare Words: a major problem for contextualized *embeddings* and how to fix it by attentive mimicking. **In Proceedings Of The Aaai Conference On Artificial Intelligence**, [S.L.], v. 34, n. 05, p. 8766-8774, 3 abr. 2020. Association for the Advancement of Artificial Intelligence (AAAI). http://dx.doi.org/10.1609/aaai.v34i05.6403

SCHOEMAKER, P. J. H. Scenario Planning: A Tool for Strategic Thinking. **Long Range Planning**, vol. 28, no. 3, p. 117, jun. 1995. https://doi.org/10.1016/0024-6301(95)91604-0

SCHOEMAKER, Paul J. H.; DAY, George S. How to make sense of weak signals. MIT Sloan Management Review, v. 50, n.3, jan. 2009. Disponível em: https://sloanreview.mit.edu/article/how-to-make-sense-of-weak-signals/

SCHOENECK, David.J.; PORTER, Alan.L.; KOSTOFF, Ronald.N.; BERGER, Elena.M. Assessment of Brazil's research literature. **Technology Analysis & Strategic Management**, v.23, N.6, p.601- 621, jul. 2011. https://doi.org/10.1080/09537325.2011.585029

SCHUMPETER, J. A. Capitalism, socialism and democracy. **The Economic Journal**, Wiley on behalf of the Royal Economic Society, v. 53, n. 212, p. 381–383, 1943.

SCHWAB, K. **The fourth Industrial Revolution**. Crown Business, New York, 2016.

SENNRICH, R.; HADDOW, B.; BIRCH, A.. Neural Machine Translation of Rare Words with Subword Units. **In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics**, Berlin, Germany, p. 1715–1725, aug. 2016. http://dx.doi.org/10.18653/v1/P16-1162

SHANNON, C. E.. A Mathematical Theory of Communication. **Bell System Technical Journal**, [*s. l*], v. 27, n. 3, p. 379-423, jul. 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

SHEN, Yung-Chi; WANG, Ming-Yeu; YANG, Ya-Chu. Discovering the Potential Opportunities of Scientific Advancement and Technological Innovation: A Case Study of Smart Health Monitoring Technology. **Technological Forecasting and Social Change**, v. 160, p. 120225, nov. 2020. https://doi.org/10.1016/j.techfore.2020.120225

SHIBATA, Naoki; KAJIKAWA, Yuya; SAKATA, Ichiro. Extracting the Commercialization Gap between Science and Technology — Case Study of a Solar Cell. **Technological Forecasting and Social Change**, v. 77, n. 7, p. 1147–55, set. 2010. https://doi.org/10.1016/j.techfore.2010.03.008

SIEGFRIED, Tom. Scientists sometimes conceal a lack of knowledge with vague words. **Science News**, 22 maio, 2020. Disponível em: https://www.sciencenews.org/article/scientists-sometimes-conceal-lack-knowledge-vague-words?

SILVA, L. S.; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. Manual de orientação. Florianópolis, 2001.

SINGH, Harjit Pal; SINGH, Sarabjeet; SINGH, J.; KHAN, S. A. VoIP: State of Art for Global Connectivity—A Critical Review. **Journal of Network and Computer Applications**, v. 37, p. 365–79, jan. 2014. https://doi.org/10.1016/j.jnca.2013.02.026

SINGH, Vikram; CHAKRABORTY, Kajal; VINCENTI, Lavina. Patent Database: Their Importance in Prior Art Documentation and Patent Search. **Journal of Intellectual Property Rights**, jan. 2016. Disponível em: http://nopr.niscair.res.in/handle/123456789/34016

SLAUGHTER, Richard A. The Foresight Principle. **Futures**, v. 22, n. 8, p. 801–19, out. 1990. https://doi.org/10.1016/0016-3287(90)90017-C

SIMON, H. A. **The Sciences of the Artificial**. 3rd ed. Cambridge: MIT Press, 1996.

SMALL, Henry; BOYACK, Kevin W.; KLAVANS, Richard. Identifying Emerging Topics in Science and Technology. **Research Policy**, v. 43, n. 8, p. 1450–67, out. 2014. https://doi.org/10.1016/j.respol.2014.02.005

SON, Hyeonju. The history of Western futures studies: an exploration of the intellectual traditions and three-phase periodization. **Futures**, [S.L.], v. 66, p. 120-137, fev. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.futures.2014.12.013

SONG, Kisik; KIM, Karp Soo; LEE, Sungjoo. Discovering New Technology Opportunities Based on Patents: Text-Mining and F-Term Analysis. **Technovation**, v. 60–61, p. 1–14, fev. 2017. https://doi.org/10.1016/j.technovation.2017.03.001

SONG, Kisik; KIM, Kyuwoong; LEE, Sungjoo. Identifying Promising Technologies Using Patents: A Retrospective Feature Analysis and a Prospective Needs Analysis on Outlier Patents. **Technological Forecasting and Social Change**, v. 128, p. 118–32, mar. 2018 https://doi.org/10.1016/j.techfore.2017.11.008

SOVIĆ, D.; BERTOŠA, B. Methods Acronyms – The Witty Side of Science, **Kemija u industriji**, v. 58, n. 7-8, p. 337–341, 2009.

STIGLER, Stephen M. Stigler's law of eponymy. **In Annals of the New York Academy of Sciences**, v.39, p. 147-157, 1980. http://dx.doi.org/10.1111/J.2164-0947.1980.TB02775.X

STUDER, R.; BENJAMINS, R. V.; FENSEL, D.: Knowledge Engineering: Principles and Methods. **Data & Knowledge Engineering**, v. 25, n. 1–2, p. 161–97, mar. 1998. https://doi.org/10.1016/S0169-023X(97)00056-6.

SU, Hsin-Ning; LEE, Pei-Chun. Mapping Knowledge Structure by Keyword Co-Occurrence: A First Look at Journal Papers in Technology Foresight. **Scientometrics**, v. 85, n. 1, p. 65–79, out. 2010. https://doi.org/10.1007/s11192-010-0259-8

SUOMINEN, Arho; NEWMAN, Nils C.. Exploring the Fundamental Conceptual Units of Technical Emergence. **In Proceedings Portland International Conference on Management of Engineering and Technology (PICMET)**, IEEE, p. 1–5, 2017. https://doi.org/10.23919/PICMET.2017.8125287

TAI, W.; KUNG, H.T.; DONG, X.. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. **Findings of the Association for Computational Linguistics**, [S.L.], 2020. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.129

TALEB, N. N.. **The black swan: The impact of the highly improbable**. Random house, vol. 2, 2007.

TAY, A. Snappy Acronyms Generate Excitement for Science (SAGES). **In The Scientist: exploring life, inspiring innovation**, fev, 2020. Disponível em: https://www.the-scientist.com/news-opinion/snappy-acronyms-generate-excitement-for-science--sages--67057

TEECE, David J.; PISANO, Gary; SHUEN, Amy. Dynamic capabilities and strategic management. **Strategic Management Journal**, [S.L.], v. 18, n. 7, p. 509-533, ago. 1997. Wiley. http://dx.doi.org/10.1002/(sici)1097-0266(199708)18:73.0.co;2-z

TECHNOLOGY FUTURES ANALYSIS METHODS WORKING GROUP (TFAMW Group). Technology Futures Analysis: Toward Integration of the Field and New Methods. **Technological Forecasting and Social Change**, v. 71, n. 3, p. 287–303, mar. 2004. https://doi.org/10.1016/j.techfore.2003.11.004

TEICHERT, T.; MITTERMAYER, M. A. Text Mining for Technology Monitoring. **IEEE International Engineering Management Conference**, v. 2, p. 596–601, 2002. https://doi.org/10.1109/IEMC.2002.1038503

THORLEUCHTER, Dirk; POEL, Dirk Van den. Weak Signal Identification with Semantic Web Mining. **Expert Systems with Applications**, v. 40, n. 12, p. 4978–85, set. 2013. https://doi.org/10.1016/j.eswa.2013.03.002

THORLEUCHTER, Dirk; POEL, Dirk Van den. Idea Mining for Web-Based Weak Signal Detection. **Futures**, v. 66, p. 25–34, fev. 2015. https://doi.org/10.1016/j.futures.2014.12.007

THORLEUCHTER, Dirk; SCHEJA, Tobias; POEL, Dirk Van den. Semantic Weak Signal Tracing. **Expert Systems with Applications**, v. 41, n. 11, p. 5009–16, set. 2014. https://doi.org/10.1016/j.eswa.2014.02.046

TIMBANE, Alexandre António. A formação de palavras a partir de siglas e acrônimos estrangeiros na língua portuguesa. **VERBUM. CADERNOS DE PÓS-GRADUAÇÃO**. ISSN 2316-3267, v. 6, p. 50-68, 2014.

TSHITOYAN, Vahe; DAGDELEN, John; WESTON, Leigh; DUNN, Alexander; RONG, Ziqin; KONONOVA, Olga; PERSSON, Kristin A.; CEDER, Gerbrand; JAIN, Anubhav. Unsupervised word *embeddings* capture latent knowledge from materials science literature. **Nature**, [S.L.], v. 571, n. 7763, p. 95-98, jul. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41586-019-1335-8

UPHAM, S. Phineas; SMALL, Henry. Emerging Research Fronts in Science and Technology: Patterns of New Knowledge Development. **Scientometrics**, v. 83, n. 1, p. 15–38, abr. 2010. https://doi.org/10.1007/s11192-009-0051-9

VAHIDNIA, Sahand; ABBASI, Alireza; ABBASS, Hussein A. Document Clustering and Labeling for Research Trend Extraction and Evolution Mapping. **In Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE at JCDL)**, p. 54-62, 2020.

VASSAKIS, Konstantinos; PETRAKIS, Emmanuel; KOPANAKIS, Ioannis. Big Data Analytics: Applications, Prospects and Challenges. **Mobile Big Data**, Springer International Publishing, v. 10, p. 3–20, 2018 https://doi.org/10.1007/978-3-319-67925-9_1

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A.N.; KAISER, L.; POLOSUKHIN, I.. Attention is all you need. **In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)**, New York-USA, p. 6000-6010, 2017.

VEEN, Barbara L. van; ORTT, J. Roland. Unifying Weak Signals Definitions to Improve Construct Understanding. Futures, v. 134, p. 102837, dez. 2021. https://doi.org/10.1016/j.futures.2021.102837

VEGA; Davide; MAGNANI, Matteo. Foundations of temporal text networks. **Applied network science**, v. 3, n. 1, p. 1-26, dez. 2018. https://doi.org/10.1007/s41109-018-0082-3

VESPIGNANI, Alessandro. Twenty Years of Network Science. **Nature**, v. 558, n. 7711, p. 528–29, jun. 2018. https://doi.org/10.1038/d41586-018-05444-y

WANG, Bin; KUO, C.-C. Jay. SBERT-WK: a sentence *embedding* method by dissecting bert-based word models. **IEEE/ACM Transactions On Audio, Speech, And Language Processing**, [S.L.], v. 28, p. 2146-2157, 2020. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/taslp.2020.3008390

WANG, Lili; LI, Zexia. Knowledge Flows from Public Science to Industrial Technologies. **The Journal of Technology Transfer**, v. 46, n. 4, p. 1232–55, ago. 2021. https://doi.org/10.1007/s10961-019-09738-9

WANG, Wei; LIU, Yun; LIU, Shuhan. Study on Assessment Model for Emerging Technology Maturity. Proccedings in Portland International Conference on Management of Engineering and Technology (PICMET), IEEE, p. 1–7, 2017. https://doi.org/10.23919/PICMET.2017.8125278

WANG, Xiaoyu; ZHAI, Yujia; LIN, Yuanhai; WANG, Fang. Mining Layered Technological Information in Scientific Papers: A Semi-Supervised Method. **Journal of Information Science**, v. 45, n. 6, p. 779–93, dez. 2019. https://doi.org/10.1177/0165551518816941

WANG, Xuefeng; QIU, Pengjun; ZHU, Donghua; MITKOVA, Liliana; LEI, Ming; PORTER, Alan L.. Identification of Technology Development Trends Based on Subject–Action–Object Analysis: The Case of Dye-Sensitized Solar Cells. **Technological Forecasting and Social Change**, v. 98, p. 24–46, set. 2015. https://doi.org/10.1016/j.techfore.2015.05.014

WANG, Xuefeng; MA, Pingping; HUANG, Ying; GUO, Junfang; ZHU, Donghua; PORTER, Alan L.; WANG, Zhinan. Combining SAO Semantic Analysis and Morphology Analysis to Identify Technology Opportunities. **Scientometrics**, v. 111, n. 1, p. 3–24 , abr. 2017. https://doi.org/10.1007/s11192-017-2260-y

WANG, Yuhu; ZHANG, Chunxia; XIANG, Shiming; PAN, Chunhong. Subgraph-Aware Graph Structure Revision for Spatial–Temporal Graph Modeling. **Neural Networks**, v. 154, p. 190–202, out. 2022. https://doi.org/10.1016/j.neunet.2022.07.017.

WANG, Zhinan; PORTER, Alan L.; WANG, Xuefeng; CARLEY, Stephen. An Approach to Identify Emergent Topics of Technological Convergence: A Case Study for 3D Printing. **Technological Forecasting and Social Change**, v. 146, p. 723–32, set. 2019. https://doi.org/10.1016/j.techfore.2018.12.015

WARNKE, P.; HEIMERIKS, G.. Technology Foresight as Innovation Policy Instrument: learning from science and technology studies. **Future-Oriented Technology Analysis**, [S.L.], p. 71-87, 2008. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-68811-2_6

WATTS, R. J.; PORTER, A. L.. Requirements-Based Knowledge Discovery for Technology Management. **Proceedings Portland International Conference on Management of Engineering and Technology - PICMET'01**, v.01, p. 71, 2001 https://doi.org/10.1109/PICMET.2001.951776

WATTS, Robert J.; PORTER, Alan L.. MINING CONFERENCE PROCEEDINGS FOR CORPORATE TECHNOLOGY KNOWLEDGE MANAGEMENT. **International Journal of Innovation and Technology Management**, v. 04, n. 02, p. 103–19, jun. 2007. https://doi.org/10.1142/S0219877007001016

WATTS, Duncan J.; STROGATZ, Steven H. Collective dynamics of 'small-world'networks. **Nature** v. 393, n. 6684, p. 440-442, 1998.

WAZLAWICK, R.S. **Metodologia de pesquisa para Ciência da Computação**. Editora Campus/Elsevier, 2014

WELLS, H.G. **Anticipations of the Reactions of Mechanical and Scientific Progress upon Human Life and Thought**. Project Gutenberg, 1902. Disponível em: http://www.gutenberg.org/etext/192291901

WELLS, Herbert G. Wanted—professors of foresight. **Futures Research Quarterly**, v. 3, n. 1, p. 89-91, 1987.

WIERINGA, Roel. Design science as nested problem solving. **In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (DESRIST '09)**. Association for Computing Machinery, New York, NY, USA, Article 8, p.1-12. 2009. https://doi.org/10.1145/1555619.1555630

WINNINK, J. J.; TIJSSEN, Robert J.W.; RAAN, A.F.J. van. Searching for New Breakthroughs in Science: How Effective Are Computerised Detection Algorithms? **Technological Forecasting and Social Change**, v. 146, p. 673–86, set. 2019. https://doi.org/10.1016/j.techfore.2018.05.018

WOLF, Thomas, *et al*. Transformers: State-of-the-Art Natural Language Processing. **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics**, p. 38–45, 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6

WU, Y. *et al*.. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016. arXiv:1609.08144

WU, F.S.; SHIU, C.C.; LEE, P.C.; SU, H.N. Integrated methodologies for mapping and forecasting science and technology trends: a case of etching technology. **Proceedings Technology Management for Global Economic Growth (PICMET)**, p. 1–23, 2010.

XU, Haiyun; WINNINK, Jos; YUE, Zenghui; LIU, Ziqiang; YUAN, Guoting. Topic-Linked Innovation Paths in Science and Technology. **Journal of Informetrics**, v. 14, n. 2, p. 101014, mai. 2020. https://doi.org/10.1016/j.joi.2020.101014

XU, Shuo; HAO, Liyuan; AN, Xin; PANG,Hongshen; LI, Ting. Review on Emerging Research Topics with Key-Route Main Path Analysis. **Scientometrics**, v. 122, n. 1, p. 607–24, jan. 2020. https://doi.org/10.1007/s11192-019-03288-5

XU, Shuo; HAO, Liyuan; YANG, Guancan; LU, Kun; AN, Xin. A Topic Models Based Framework for Detecting and Forecasting Emerging Technologies. **Technological Forecasting and Social Change**, v. 162, p. 120366, jan. 2021. https://doi.org/10.1016/j.techfore.2020.120366

XU, Shuo; ZHAI, Dongsheng; WANG, Feifei; AN, Xin; PANG, Hongshen; SUN, Yirong. A Novel Method for Topic Linkages between Scientific Publications and Patents: A Novel Method for Topic Linkages between Scientific Publications and Patents. **Journal of the Association for Information Science and Technology**, v. 70, n. 9, p. 1026–42, set. 2019. https://doi.org/10.1002/asi.24175

YANG, Bo; LIU, Dayou; LIU, Jiming. Discovering Communities from Social Networks: Methodologies and Applications. **Handbook of Social Network Technologies and Applications**, Springer US, p. 331–46, 2010. https://doi.org/10.1007/978-1-4419-7142-5_16

YANG, Heyoung; HA, Taehyun; HONG, Sunghwa; KIM, Kyuli. **Emerging Weak Signals 2023 in Science and Technology**, KISTI Data Insight Report no.24, ISBN 978-89-294-1341-5-93500, 2022.

YANG, Jaewon; LESKOVEC, Jure. Defining and Evaluating Network Communities Based on Ground-Truth. **Knowledge and Information Systems**, v. 42, n. 1, p. 181–213, jan. 2015. https://doi.org/10.1007/s10115-013-0693-z.

YANG, Junyong; ZHONG, Ming; ZHU, Yuanyuan; QIAN, Tieyun; LIU, Mengchi; YU, Jeffrey Xu. Scalable Time-Range k -Core Query on Temporal Graphs. **Proceedings of the VLDB Endowment**, v. 16, n. 5, p. 1168–80, jan. 2023. https://doi.org/10.14778/3579075.3579089

YAU, Chyi-Kwei; PORTER, Alan L., NEWMAN, Nils; SUOMINEN, Arho. Clustering Scientific Documents with Topic Modeling. **Scientometrics**, v. 100, n. 3, p. 767–86, set. 2014. https://doi.org/10.1007/s11192-014-1321-8

YOO, Sun Hi; WON, DongKyu. Simulation of Weak Signals of Nanotechnology Innovation in Complex System. **Sustainability**, v. 10, n. 2, p. 486, fev. 2018. https://doi.org/10.3390/su10020486

YOON, Janghyeok. Detecting Weak Signals for Long-Term Business Opportunities Using Text Mining of Web News. **Expert Systems with Applications**, v. 39, n. 16, p. 12543–50, nov. 2012. https://doi.org/10.1016/j.eswa.2012.04.059

YOON, Byungun; LEE, Sungjoo. Applicability of Patent Information in Technological Forecasting: A Sector-specific Approach. **Journal of Intellectual Property Rights**, v.l 17, p 37-45, jan. 2012.

ZHAN, Chuan; DU, Ye. Early Identification Methods for Emerging Technologies Based on Weak Signals. **preprint, In Review**, 2022. https://doi.org/10.21203/rs.3.rs-2291140/v1

ZHANG, Y.; ROBINSON, D.; PORTER, A.L.; ZHU, D.; ZHANG, G., LU, J.. Technology roadmapping for competitive technical Intelligence. **Technological Forecasting and Social Change**, v. 110, p. 175–86, set. 2016a. https://doi.org/10.1016/j.techfore.2015.11.029

ZHANG, Yi, ZHANG, Guangquan; CHEN, Hongshu; PORTER, Alan L.; ZHU, Donghua; LU, Jie. Topic Analysis and Forecasting for Science, Technology and Innovation: Methodology with a Case Study Focusing on Big Data Research. **Technological Forecasting and Social Change**, v. 105, p. 179–91, abr. 2016b. https://doi.org/10.1016/j.techfore.2016.01.015

ZHANG, Yi; ZHANG, Guangquan; ZHU, Donghua; LU, Jie. Scientific Evolutionary Pathways: Identifying and Visualizing Relationships for Scientific Topics. **Journal of the Association for Information Science and Technology**, v. 68, n. 8, p. 1925–39, ago. 2017. https://doi.org/10.1002/asi.23814

ZHANG, Yi; ZHOU, Xiao; PORTER, Alan L.; GOMILA, Jose M. Vicente. How to Combine Term Clumping and Technology Roadmapping for Newly Emerging Science & Technology Competitive Intelligence: 'Problem & Solution' Pattern Based Semantic TRIZ Tool and Case Study. **Scientometrics**, v. 101, n. 2, p. 1375–89, nov. 2014. https://doi.org/10.1007/s11192-014-1262-2

ZHANG, Yifei; LIN, Long-Long; YUAN, Ping-Peng; JIN, Hai. Significant Engagement Community Search on Temporal Networks: Concepts and Algorithms. 2022. https://doi.org/10.48550/ARXIV.2206.06350.

ZHAO, Dongyuan; TANG, Zhongjun; HE, Duokui. A systematic literature review of weak signal identification and evolution for corporate foresight, **Kybernetes**, v. ahead-of-print, n. Ahead-of-print, 2023. https://doi.org/10.1108/K-03-2023-0343

ZHAO, Wayne Xin, *et al*. A Survey of Large Language Models. 2023. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.2303.18223.

ZHAO, Yifeng; WANG, Xiangwei; YANG, Hongxia; SONG, Le; TANG, Jie. Large Scale Evolving Graphs with Burst Detection. **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence**, p. 4412–18, 2019. https://doi.org/10.24963/ijcai.2019/613

ZHOU, Yuan; DONG, Fang; KONG, Dejing; LIU, Yufei. Unfolding the Convergence Process of Scientific Knowledge for the Early Identification of Emerging Technologies. **Technological Forecasting and Social Change**, v. 144, p. 205–20, jul. 2019. https://doi.org/10.1016/j.techfore.2019.03.014

ZHOU, Yuan; DONG, Fang; LIU, Yufei; LI, Zhaofu; DU, JunFei; ZHANG, Li. Forecasting Emerging Technologies Using Data Augmentation and Deep Learning. **Scientometrics**, v. 123, n. 1, p. 1–29, abr. 2020. https://doi.org/10.1007/s11192-020-03351-6

ZHOU, Xiao; HUANG, Lu; ZHANG, Yi; YU, Miao. A Hybrid Approach to Detecting Technological Recombination Based on Text Mining and Patent Network Analysis. **Scientometrics**, v. 121, n. 2, p. 699–737, nov. 2019. https://doi.org/10.1007/s11192-019-03218-5

ZHU, Chun-Xue; LIN, Long-Long; YUAN, Ping-Peng; JIN, Hai. Discovering Cohesive Temporal Subgraphs with Temporal Density Aware Exploration. **Journal of Computer Science and Technology**, v. 37, n. 5, p. 1068–85, out. 2022. https://doi.org/10.1007/s11390-022-2431-z.

ZHU, Donghua; PORTER, Alan L. Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting. **Technological Forecasting and Social Change**, v. 69, n. 5, p. 495–506, jun. 2002. https://doi.org/10.1016/S0040-1625(01)00157-3

ZHU, Lin; ZHU, Donghua; WANG, Xuefeng; CUNNINGHAM, Scott W.; WANG, Zhinan. An Integrated Solution for Detecting Rising Technology Stars in Co-Inventor Networks. **Scientometrics**, v. 121, n. 1, p. 137–72, out. 2019. https://doi.org/10.1007/s11192-019-03194-w

ZIPF, George Kingsley. **Human behavior and the principle of least effort: An introduction to human ecology**, Addison-Wesley, Cambridge, MA, 1949.