UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS ARARANGUÁ
CENTRO DE CIÊNCIAS, TECNOLOGIA E SAÚDE (CTS)
TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

Pollyanna Cardoso Souza

Desvendando Padrões e Insights: Visualização de Ideias com Modelos de Linguagem e Análise de Agrupamento

Araranguá
2024

Pollyanna Cardoso Souza

Desvendando Padrões e Insights: Visualização de Ideias com Modelos de Linguagem e Análise de Agrupamento

Trabalho de Conclusão de Curso submetido ao curso de Tecnologias da Informação e Comunicação do Centro de Ciências, Tecnologia e Saúde do Campus de Araranguá da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel(a) em Tecnologias da Informação e Comunicação.

Orientador(a): Prof.(a) Dr.(a) Marina Carradore Sérgio,

Araranguá
2024

Souza, Pollyanna Cardoso
Desvendando Padrões e Insights: Visualização de Ideias com Modelos de Linguagem e Análise de Agrupamento / Pollyanna Cardoso Souza ; orientadora, Marina Carradore Sérgio, 2024.
86 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2024.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Gestão de Ideias. 3. Inovação. 4. Busca Semântica. 5. Modelos de Linguagem. I. Sérgio, Marina Carradore. II. Universidade Federal de Santa Catarina. Graduação em Tecnologias da Informação e Comunicação. III. Título.

Pollyanna Cardoso Souza

Desvendando Padrões e Insights: Visualização de Ideias com Modelos de Linguagem e Análise de Agrupamento

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de bacharel e aprovado em sua forma final pelo Curso de Tecnologias da Informação e Comunicação.

Araranguá, 11 de dezembro de 2024.

Documento assinado digitalmente
Fernando Jose Spanhol
Data: 20/12/2024 15:40:16-0300
CPF: ***.656.419-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Fernando José Spanhol, Dr.
Coordenação do Curso

Banca examinadora

Documento assinado digitalmente
Marina Carradore Sergio
Data: 19/12/2024 20:33:19-0300
CPF: ***.746.359-**
Verifique as assinaturas em https://v.ufsc.br

Profª. Marina Carradore Sérgio, Drª.
Orientadora

Documento assinado digitalmente
Alexandre Leopoldo Goncalves
Data: 19/12/2024 22:10:21-0300
CPF: ***.740.349-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Alexandre Leopoldo Gonçalves, Dr.
Universidade de Federal de Santa Catarina

Documento assinado digitalmente
Cristian Cechinel
Data: 19/12/2024 23:23:16-0300
CPF: ***.628.449-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Cristian Cechinel, Dr.
Universidade de Federal de Santa Catarina

Araranguá, 2024.

AGRADECIMENTOS

Agradeço primeiramente à minha orientadora, Profª Dr. Marina Carradore, por ter sido excepcionalmente atenciosa e solícita durante todo o processo de desenvolvimento deste trabalho. Orientação esta que levarei para a vida, pois abrange muito mais que somente um trabalho acadêmico. Seu exemplo e força de vontade se tornaram uma grande inspiração para mim.

Agradeço ainda aos docentes que participaram do meu processo de aprendizado durante a graduação, pois atender as aulas destes e escutar as experiências de cada um, fizeram parte do meu processo de amadurecimento como estudante.

Por fim, agradeço a banca avaliadora, aos amigos e familiares envolvidos.

RESUMO

A inovação desempenha um papel fundamental no desenvolvimento estratégico das organizações, sendo frequentemente impulsionada por ideias que emergem de diversas fontes. As plataformas de gestão de ideias surgem como ferramentas para coletar ideias e sugestões, permitindo explorar a criatividade coletiva e identificar oportunidades estratégicas. Este trabalho tem como objetivo analisar ideias e sugestões coletadas de um sistema de gestão de ideias, com foco na identificação de padrões temáticos e tendências úteis para a gestão estratégica. Para alcançar esse objetivo, foi utilizada uma abordagem metodológica integrada, combinando vetorização de texto com SBERT, agrupamento com K-means, redução de dimensionalidade com PCA, e busca semântica com modelos de linguagem (LLM). Os resultados revelaram que, embora as ideias estivessem distribuídas em diferentes categorias e agrupamentos, emergiram padrões temáticos recorrentes, conectando áreas distintas. Entre os destaques, identificou-se o uso de tecnologias sustentáveis como um tema transversal e inovações significativas na área da saúde. O método desenvolvido provou ser eficiente na extração de insights relevantes, contribuindo para apoiar gestores e empreendedores na identificação de oportunidades de mercado e no aprimoramento de produtos e processos. A análise também mostrou que a combinação de tecnologias emergentes com abordagens sustentáveis tem um grande potencial para gerar soluções inovadoras. Este estudo contribui para a compreensão de como a análise de grandes volumes de dados pode ser utilizada para a gestão estratégica de ideias.

Palavras-chave: Gestão de Ideias; Inovação; Busca Semântica; Modelos de Linguagem.

ABSTRACT

Innovation plays a key role in the strategic development of organizations, often driven by ideas emerging from various sources. Idea management platforms have emerged as tools to collect ideas and suggestions, enabling the exploration of collective creativity and the identification of strategic opportunities. This paper aims to analyze ideas and suggestions collected from an idea management system, focusing on the identification of thematic patterns and trends useful for the strategic management. To achieve this objective, an integrated methodological approach was used, combining text vectorization with SBERT, clustering with K-means, dimensionality reduction with PCA, and semantic search with language models (LLM). The results revealed that, although ideas were distributed across different categories and clusters, recurring thematic patterns emerged, connecting distinct areas. Notably, the use of sustainable technologies was identified as a cross-cutting theme, along with significant innovations in the healthcare sector. The developed method proved effective in extracting relevant insights, helping managers and entrepreneurs identify market opportunities and improve products and processes. The analysis also showed that combining emerging technologies with sustainable approaches has great potential to generate innovative solutions. This study contributes to the understanding of how large volumes of data analysis can be used for strategic idea management.

Keywords: Idea Management; Innovation; Semantic Search; Language Models.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BERT - *Bidirectional Encoder Representations from Transformers*

DSR - *Design Science Research*

IA - Inteligência Artificial

ILN - Interpretação de Linguagem Natural

KMS - *Knowledge management systems*

LLM - *Large Language Model*

NLG - *Natural Language Generation*

NLP - *Natural Language Processing*

NLR - *Natural Language Reasoning*

PCA - *Principal Components Analysis*

SGC - Sistemas de Gerenciamento de Conhecimento

SGI - Sistemas de Gestão de Ideias

TF-IDF - *Term Frequency-Inverse Document Frequency*

SUMÁRIO

1 INTRODUÇÃO

No cenário empresarial atual, a gestão eficaz de ideias tornou-se um pilar essencial para o desenvolvimento e a inovação contínua das empresas. Valorizar as ideias permite que as organizações não apenas se mantenham competitivas, mas também liderem na criação de soluções inovadoras. De acordo com Mikesone et al. (2020), o gerenciamento de ideias envolve a identificação, geração e avaliação de ideias, destacando sua importância para o crescimento e a sustentabilidade organizacional. A capacidade de transformar ideias em valor está diretamente correlacionada ao sucesso a médio e longo prazo de qualquer empresa (Mikesone et al., 2024).

Vierula (2024) reforça essa visão ao afirmar que a gestão de ideias é crucial na era da informação, onde a inovação oferece uma vantagem competitiva significativa. Rampa e Agogué (2021) argumentam que ideias bem geridas podem se transformar em inovações disruptivas que reposicionam uma empresa no mercado. Nesse contexto, a Gestão de Ideias não é apenas um facilitador, mas uma peça fundamental no processo de inovação, como destacado por Barbieri, Álvares e Cajazeira (2009) e Cooper, Edgett e Kleinschmidt (2001).

Ao longo dos anos, foram desenvolvidos diversos modelos e ferramentas para facilitar o processamento sistemático de ideias, com o objetivo de extrair insights valiosos que possam gerar inovação (Ahmed, 2008; Bakker, Boersma e O'Reel, 2006; Brem e Voigt, 2009). A gestão de ideias deve ser coordenada, conectando ideias semelhantes para gerar valor e inovação dentro das organizações (Thom, 2003).

Com o advento da tecnologia, o processamento manual de grandes volumes de ideias tornou-se inviável. Ferramentas automatizadas, como as baseadas em Processamento de Linguagem Natural (NLP - *Natural Language Processing*), surgiram como soluções eficazes para otimizar a gestão de ideias. O NLP, fundamentado em inteligência artificial, permite a análise de grandes volumes de texto, identificando padrões e gerando *insights* relevantes (Liddy, 2001; Chowdhury, 2003). Mais recentemente, técnicas avançadas como aprendizado de máquina, aprendizado profundo e modelos de linguagem de larga escala (LLM) têm sido incorporadas aos Sistemas de Gestão de Ideias (SGI) para melhorar ainda mais a eficácia desse processo (Naveed et al., 2023).

Aprendizado de máquina, do inglês *Machine Learning*, é definido por Mitchell (1997) como o estudo de algoritmos que melhoram automaticamente através da experiência. Já o Aprendizado Profundo, do inglês *Deep Learning*, utiliza redes neurais profundas e tem se mostrado particularmente eficaz na extração de informações relevantes de grandes volumes de dados (Shinde; Shah, 2018). Recentemente, modelos de linguagem pré-treinados aplicados em larga escala, como o BERT, têm se difundido e alcançado resultados de nível humano em tarefas relacionadas à linguagem (Kasneci et al., 2023).

A aplicação dessas tecnologias nos Sistemas de Gestão de Ideias permite não apenas a organização e avaliação de ideias, mas também a extração de *insights* como emoções, opiniões e tendências (Berger; Packard, 2022). Dessa forma, as empresas podem aproveitar de maneira mais eficaz o vasto conjunto de dados não estruturados disponíveis, transformando ideias em inovação.

Nesse contexto, este trabalho busca analisar as ideias coletadas de um site de geração de ideias utilizando algoritmos de vetorização como o modelo SBERT, técnicas de agrupamento, como o K-means, e técnicas de busca semântica baseadas em LLM, com o objetivo de identificar padrões e *insights* que possam aprimorar os processos de gestão de ideias em organizações.

1 PROBLEMÁTICA

No âmbito da gestão de ideias nas organizações, há uma lacuna no investimento em pesquisas que apontem melhorias e benefícios da intersecção entre inteligência artificial e *softwares* de gestão de ideias, apesar de o tema de Gestão de Ideias ser objeto de estudos desde 1982 (Serena, 2024). Mikelsone e Lielã (2015) identificaram, através de uma revisão sistemática da literatura, 18 tópicos potenciais de pesquisa relacionados ao tema, sugerindo a realização de uma pesquisa holística sobre os fatores de sucesso na aplicação de sistemas de gestão de ideias.

Brem, Giones e Werle (2023) questionam como a inteligência artificial poderia acelerar os processos de gestão da inovação, promovendo maior agilidade na concepção e desenvolvimento de novos produtos e processos. Mariani, Machado e

Nambisan (2022) questionam o impacto do nível de inovação dos processos em uma empresa na adoção da IA.

Neste contexto, a questão central é: como a aplicação de técnicas de processamento de linguagem natural com algoritmos de LLM poderia extrair insights, tendências e temas relevantes de uma base de ideias?

1.2 OBJETIVOS

Para melhor compreensão dos objetivos deste projeto, estes foram divididos em objetivo geral e objetivos específicos, delimitados a seguir.

1.2.1 Objetivo Geral

Desenvolver um método que integre algoritmos de agrupamento e modelos de linguagem para analisar ideias coletadas de sites especializados, com foco na identificação de oportunidades, padrões e tendências.

1.2.2 Objetivos Específicos

- Coletar ideias de sites especializados, identificando a natureza das ideias coletadas;
- Integrar algoritmos de agrupamento com Modelos de Linguagem de Larga Escala (LLMs), para realizar a análise das ideias coletadas;
- Disponibilizar resultados que contenham oportunidades, padrões e tendências emergentes nas ideias analisadas.

1.3 JUSTIFICATIVA

Em um contexto empresarial altamente competitivo, as organizações que conseguem gerar ideias inovadoras possuem uma vantagem significativa sobre seus concorrentes (Gerlach e Brem, 2017). Para alcançar esse diferencial competitivo e atender às demandas dos *stakeholders*, é essencial adotar métodos eficazes para a obtenção de *insights* que impulsionem a inovação em produtos e processos (Lawson; Samson, 2001).

Dados textuais representam uma fonte rica de informações, incluindo ideias e sugestões inovadoras que muitas vezes estão dispersas em diversas plataformas, como redes sociais, *blogs* e sites especializados, como o "CoolBusinessIdeas.com". No entanto, a análise manual desses dados em larga escala é inviável devido ao alto custo e ao tempo necessário para o processamento (Zhai; Massung, 2016).

Com os avanços na capacidade de *hardware* e no desenvolvimento de inteligência artificial, especialmente no campo do processamento de linguagem natural (NLP), a aplicação de algoritmos de aprendizado de máquina, como os modelos de linguagem de larga escala (LLMs), tornou-se não apenas viável, mas essencial para a extração eficiente de ideias a partir de textos (Naveed et al., 2023). A rápida adoção dessas tecnologias pelas empresas reflete sua eficácia em transformar grandes volumes de dados textuais em *insights* acionáveis.

Nesse contexto, este projeto se justifica pela necessidade de empregar técnicas avançadas de processamento de linguagem natural e aprendizado de máquina para extrair *insights* dos vastos volumes de dados textuais das ideias disponíveis. Essa abordagem permitirá uma análise mais rápida, eficiente e abrangente das ideias e sugestões apresentadas, potencializando a capacidade de inovação das empresas.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está estruturado em cinco capítulos, conforme descrito a seguir:

- Capítulo 1: Introdução — Apresenta a temática do trabalho, os objetivos gerais e específicos, a justificativa e a problemática do estudo;
- Capítulo 2: Fundamentação Teórica — Aborda os principais conceitos relacionados à gestão de ideias, sistemas de gestão de ideias, processamento de linguagem natural, modelos de linguagem de larga escala (LLMs) e *embeddings*, além de trabalhos correlatos;
- Capítulo 3: Metodologia — Descreve a metodologia adotada para a execução do trabalho, incluindo os métodos e ferramentas utilizados;
- Capítulo 4: Apresentação e discussão dos Resultados — Apresenta e discute os resultados obtidos, incluindo discussões sobre os padrões identificados e as possíveis direções para pesquisas futuras;
- Capítulo 5: Considerações Finais — Reflete sobre os resultados alcançados, e as contribuições do estudo.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 GESTÃO DE IDEIAS

2.1.1 Ideia

A ideia surge do pensamento humano e se manifesta de várias formas, como em opiniões, imagens, pesquisas, planos, previsões e avaliações sobre diversos assuntos (Lee, 2012). Michalko (2003) complementa, afirmando que a ideia pode ser vista como um pensamento produtivo, que busca novas alternativas para problemas já conhecidos.

Tidd, Bessant e Pavitt (2005) apontam que as ideias são conceitos ou pensamentos que se tornam tangíveis, resultando em invenções. Gurteen (1998) define a ideia como algo ainda não testado, comprovado ou aplicado, enquanto Koen et al. (2002) a consideram a solução primária para um problema já elaborado.

2.1.2 Inovação

A inovação, dentro de um contexto organizacional, refere-se à capacidade da empresa de oferecer novidades ou melhorias em seus produtos ou serviços que agreguem valor para seus *stakeholders*, clientes ou sociedade (BAREGHEH et al., 2009). Além disso, pode ser entendida como "uma oportunidade de criar valor para novos investimentos" (Kornish; Ulrich, 2014, p.15).

As inovações representam um processo disruptivo de aplicação de ideias por meio de abordagens nunca antes exploradas, em produtos e processos existentes, visando gerar valor e adaptabilidade diante das mudanças externas enfrentadas pelas empresas (Mckeown, 2008; Kang; Kang, 2009).

A gestão de ideias, como parte essencial da gestão da inovação, considera que as ideias que passam pelos processos de seleção e julgamento adequados, quando implementadas com sucesso, podem ser consideradas inovadoras. Isso resulta na melhoria de produtos e processos, bem como na descoberta de invenções, sendo um aspecto fundamental no processo de inovação das organizações (Thorleuchter; Van Den Poel, 2016; Gerlach; Brem 2017).

2.1.3 Gestão de Ideias

A gestão de ideias pode ser compreendida como um subprocesso da gestão de inovação, que possui como objetivo principal a coleta, desenvolvimento ou geração, avaliação e seleção das melhores e mais eficazes ideias, a fim de criar ou melhorar produtos e serviços (Brem; Voigt, 2007; Saldivar et al. 2019; Gochermann e Nee, 2019).

Dentro das organizações, a gestão de ideias pode ser utilizada com a finalidade de promover inovações, trazendo assim, diversos benefícios, como a otimização dos processos realizados na empresa, aumento na produtividade organizacional, criação não só e de novos produtos e serviços, mas também está relacionado com desenvolvimento de novos processos internos, estratégia e *mindset* (Brem; Voigt, 2007), diminuição dos custos de investimentos e vantagem competitiva organizacional e a criação de valor por meio de ideias (Sint et al., 2010; Xie; Zhang, 2010; Westerski; Iglesias, 2011; Poveda; Westerski; Iglesias, 2012; Karimi-majd; Mahootchi, 2015).

Green et al (1983) anteviu que o gerenciamento de ideias poderia vir a tornar-se um grande aliado na gestão em si, percebendo que o processo de confecção, captura e análise poderia ser feito por intermédio tecnológico.

De acordo com Bakker (2010), as organizações que adotam uma gestão de ideias eficaz podem colher diversos resultados positivos, incluindo redução do tempo, aumento da eficiência, melhoria da rentabilidade, maior agilidade e diminuição dos custos (Bakker, 2010). Gibson e Skarzynski (2008) e Bakker (2010) destacam ainda que compreender os processos criativos e a dinâmica de geração de novas ideias é fundamental para estabelecer inovações auto sustentáveis.

2.2 SISTEMAS DE GESTÃO DE IDEIAS

Os sistemas de gestão de ideias que hoje se encontram difundidos entre as organizações, são uma resposta tecnológica às antigas caixas de sugestões (Turrell, 2002). Lin (2022) pontua que Sistemas de Gerenciamento de Conhecimento (KMS - *Knowledge management systems*) ou ainda, Sistemas de Gestão de Ideias (SGI), são as engrenagens da era da informação, acentuando ainda que o método de

extração de conhecimento a partir dos dados inseridos em tais sistemas podem melhorar a compreensão das informações obtidas, colaboração entre as pessoas e alinhamento dos processos realizados dentro das organizações.

Martinez-Torres e Olmedilla (2016), Gabriel et al. (2016b) e Dziallas e Blind (2019) apontam que por conta da alta volumetria de dados não estruturados que são submetidos nos SGI, torna-se difícil para que analistas das organizações façam o processamento e análise destes dados manualmente. A tecnologia, então, foi adotada para otimizar esse processo dentro da gestão de ideias. Stenmark (2000) observa que a substituição dos modelos físicos de Gestão de Ideias por *softwares* oferece muitas vantagens.

Segundo Westerski, Dalamagas e Iglesias (2012) as últimas décadas trouxeram uma evolução no que diz respeito à cobertura destes sistemas para as empresas. Antes o que somente recebia sugestões internas a organização, agora conta com a internet como aliada, transformando os *stakeholders* em agentes de ideação, bem como uma melhora colaborativa das ideias e sinergia entre estas e os outros processos que a envolvem.

Segundo Leka (2024), houve uma recente integração de inteligência artificial, como os algoritmos de aprendizado de máquina, aprendizado profundo e processamento de linguagem natural nos *softwares* de gerenciamento de ideias. Brem, Giones e Werle (2021) compilaram os benefícios e usabilidades dos diversos aspectos da IA dentro do contexto de gestão, como por exemplo, sendo um facilitador dos processos de inovação.

A integração de IA nos *softwares* de gerenciamento de ideias pode aumentar a capacidade de inovação (Mikelsone; Uvarova; Seger, 2022), motivar os funcionários a gerarem novas ideias (Ekman; Dahlin, 2011), e construir vantagens competitivas no mercado (Kabir, 2017). Segundo Hu e Xu (2023), os SGI têm se tornado um dos catalisadores responsáveis pelo aumento das vendas de novos produtos no mercado.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Devido à importância da linguagem na comunicação e interações humanas, a demanda por mecanismos capazes de lidar com tarefas linguísticas complexas,

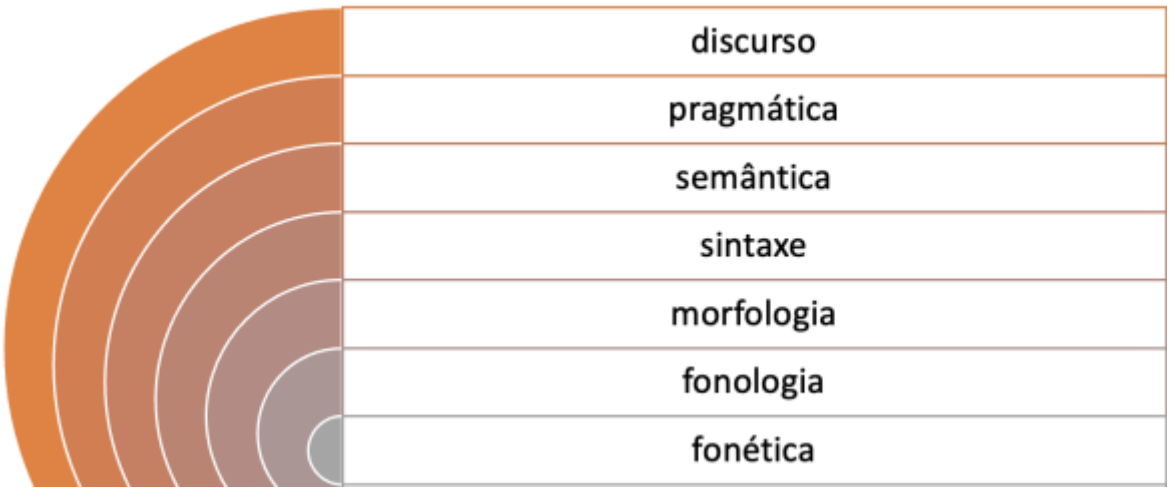
como tradução em diferentes idiomas, sumarização de textos, extração de informações e compreensão, tem aumentado constantemente (Naveed et al., 2023).

O Processamento de Linguagem Natural (NLP) tem sido utilizado desde meados de 1940, e primeiramente surgiu com intuito de se assimilar com o processamento humano de textos, como em conversões idiomáticas, sumarização textual, entre outros (Hirschberg, Manning, 2017; Manning, Socher, 2019; Kulkarni, Shivananda, 2019).

Segundo Caseli, Nunes e Pagano (2023), o NLP não se limita apenas a texto, abrangendo a língua falada (voz) e até mesmo a linguagem de sinais já tem sido alvo de pesquisa na área, sendo cada aspecto deste, tratado computacionalmente pelo NLP de um modo diferente.

Como esquematizado na Figura 1, do centro para as bordas, o NLP abrange atualmente todo espectro da fala, sendo textual ou verbal, desde a organização sonora dos morfemas para criação de frases e sentenças, até a significação de cada palavra dentro de um contexto.

Figura 1 - Abrangência do NLP



Fonte: Caseli; Nunes; Pagano, 2023.

Ainda, o NLP se divide em duas grandes subáreas de estudo: Interpretação de Linguagem Natural (NLI) e Geração de Linguagem Natural (NLG). A primeira diz respeito a tudo aquilo que envolve segmentação e classificação dos componentes linguísticos e interpretação da mesma. Já a segunda, diz respeito a geração de

linguagem, como os chatbots, onde busca-se construir um diálogo, com respostas de nível humano, como exemplo o Chat GPT.

Frequentemente, técnicas como *machine learning* tem sido empregadas conjuntamente com NLP para obtenção de melhores resultados (Shinde; Shah, 2018)

2.3.1 Modelos de Linguagem de Grande Escala (LLMs)

Os avanços tecnológicos em *hardware* na última década possibilitaram o desenvolvimento significativo dos modelos linguísticos de grande escala (LLMs). Esses avanços permitiram treinamentos em larga escala, resultando em modelos que competem com o desempenho humano em diversas tarefas (Wang; Pruksachatkun, 2019; Adiwardana et al., 2020). Arcas (2022) destaca que os LLMs foram projetados para serem disruptivos em sua performance, processando e gerando textos com uma coerência notável.

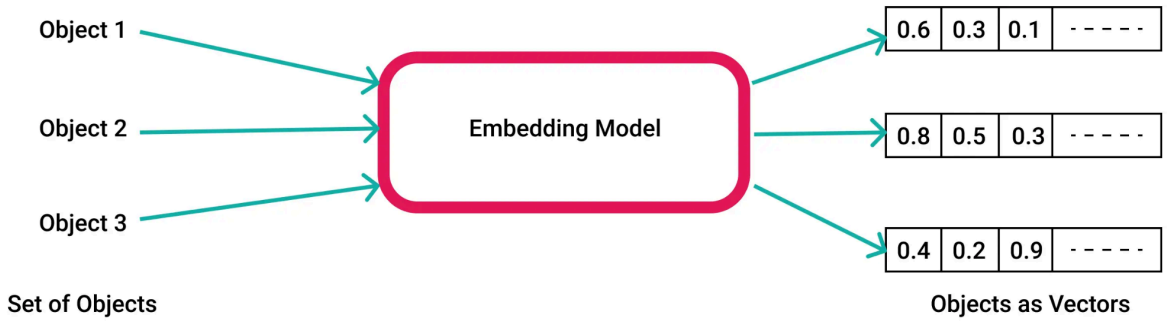
Diferentemente dos métodos tradicionais de processamento de linguagem, que são orientados a tarefas específicas e supervisionadas, os modelos linguísticos pré-treinados utilizam treinamento auto-supervisionado em grandes corpos textuais (Naveed et al., 2023). Esses modelos internalizam representações textuais genéricas, o que lhes permite alcançar resultados impressionantes. Modelos como GPT-3, GLaM, AlphaCode e Bard exemplificam os avanços significativos na área (Naveed et al., 2023; Petukhova, 2024).

2.3.2 *Embeddings*

O conceito de *embeddings* refere-se ao processo de vetorização de palavras, no qual cada palavra é atribuída a um vetor com valor real que reflete seu valor semântico. Na esquematização apresentada na Figura 2, pode-se analisar o conceito do processo de vetorização. Pennington, Socher e Manning (2014) definem *embeddings* como uma técnica para representar palavras em um espaço vetorial, facilitando a análise e a comparação semântica. Smilkov et al. (2016) complementam afirmando que o *embedding* organiza palavras em um espaço

euclidiano *n*-dimensional, onde a proximidade entre vetores reflete a similaridade semântica.

Figura 2 - Processo conceitual de vetorização

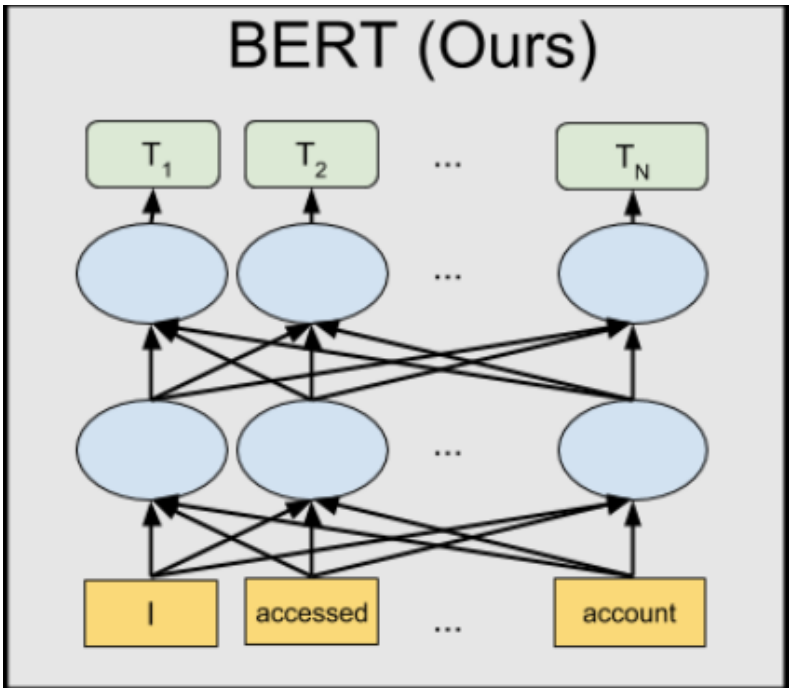


Fonte: Pinecone (2024).

A importância dos *embeddings* reside em sua capacidade de organizar grandes volumes de dados desestruturados em categorias significativas, permitindo a extração de informações valiosas e uma melhor compreensão dos temas abordados (Petukhova, 2024). Técnicas de vetorização textual, como *Term Frequency-Inverse Document Frequency* (TF-IDF) (Salton, 1983), *Word2Vec* (Mikolov et al., 2013) e *GloVe* (Pennington; Socher, 2014), embora eficazes, apresentavam limitações devido à representação estática de palavras com múltiplos significados.

A introdução do método BERT (*Bidirectional Encoder Representations from Transformers*) marcou um avanço significativo, permitindo que uma mesma palavra tenha diferentes vetores dependendo do seu contexto (Devlin et al., 2019), como mostrado na Figura 3. Isso revolucionou a representação semântica, tornando-a mais rica e adaptável às nuances do significado das palavras (Petukhova, 2024).

Figura 3 - Representação da arquitetura neural BERT



Fonte: Adaptado de Google Research (2018)

O algoritmo BERT possui como diferencial a possibilidade de uma mesma palavra ter mais de um valor semântico, a depender do contexto que esta se encontra, isso porque este algoritmo é pré-treinado em um grande *set* de treinamento, como *Semi-supervised Sequence Learning, Generative Pre-Training*, ELMo, e ULMFit (Devlin; Chang, 2018). Assim, ao rodar em cima de um conjunto textual (retângulos amarelos), este faz o batimento contextual semântico entre as diversas camadas da sua rede neural (setas percorrendo os balões azuis), atribuindo ao final um valor a cada palavra (retângulo verde “T”).

2.4 ALGORITMOS DE AGRUPAMENTO

Os algoritmos de agrupamento, ou *clustering*, são ferramentas fundamentais na análise de dados, permitindo a segmentação de um conjunto de dados em grupos ou *clusters* de objetos semelhantes. Diferentes algoritmos foram desenvolvidos para atender a várias necessidades e características dos dados.

Cassiano e Souza (2014), Agrawal et al. (1998), Ester et al. (1996), Ng e Han (1994), e Han e Kamber (2001) destacaram os requisitos ideais que um algoritmo de agrupamento deve atender:

a) Descobrir *clusters* com formatos diversos: Os *clusters* podem ter formas variadas no espaço euclidiano, como esféricas, lineares, alongadas, elípticas, cilíndricas, espirais, entre outras;

b) Identificar *clusters* de diferentes tamanhos: O algoritmo deve ser capaz de detectar grupos com tamanhos variados;

c) Aceitar diferentes tipos de variáveis: Os métodos devem lidar com variáveis de vários tipos, como intervalares, binárias, categóricas, ordinais, proporcionais, ou combinações dessas;

d) Não ser afetado pela ordem de apresentação dos objetos: A ordem em que os objetos são apresentados não deve alterar os resultados, garantindo consistência mesmo com diferentes sequências de entrada;

e) Suportar objetos com qualquer número de atributos (dimensões): Enquanto a visualização humana é limitada a três dimensões, os métodos devem processar objetos de alta dimensionalidade e gerar resultados compreensíveis mesmo sem uma visualização direta;

f) Ser escalável para grandes volumes de dados: O método deve lidar com grandes conjuntos de dados, possivelmente contendo milhões de objetos, com eficiência e rapidez, independentemente da quantidade de dimensões ou objetos;

g) Gerar resultados interpretáveis e úteis: Os *clusters* devem ser descritos de forma simples e compreensível para que os usuários possam entender e utilizar os resultados de maneira prática;

h) Ser resistente à presença de ruído: Muitos conjuntos de dados reais possuem ruídos ou dados incompletos, desconhecidos ou incorretos, e a qualidade dos *clusters* não deve ser comprometida por esses problemas;

i) Exigir o mínimo de conhecimento sobre os parâmetros de entrada: Como os parâmetros apropriados nem sempre são claros, especialmente em grandes conjuntos de dados, os métodos devem funcionar bem mesmo quando há pouca informação disponível sobre esses parâmetros;

j) Aceitar restrições: Em muitas aplicações práticas, é necessário agrupar os dados conforme restrições específicas, e o método deve ser capaz de formar *clusters* que respeitem essas limitações;

k) Determinar o número adequado de *clusters*: Descobrir a quantidade ideal de *clusters* em um conjunto de dados é desafiador, e muitos métodos exigem um número de referência para definir isso;

No entanto, segundo Agrawal et al. (1998), nenhum algoritmo atende a todos esses requisitos de forma perfeita. Portanto, diversas abordagens foram desenvolvidas para lidar com uma variedade de casos (Cassiano; Souza, 2014).

Pauletic, Prskalo e Bakaric (2019) definem seis categorias principais de algoritmos de agrupamento, cada uma adequada para diferentes características dos dados e objetivos de análise:

- Método Hierárquico: Os algoritmos hierárquicos criam uma estrutura de árvore para os dados, podendo seguir uma abordagem aglomerativa (*bottom-up*), que começa com cada objeto como um *cluster* individual e mescla os *clusters* mais próximos, ou uma abordagem divisiva (*top-down*), que começa com todos os objetos em um único *cluster* e divide-o até que cada objeto seja um *cluster* por si só;
- Método Baseado em Densidade: Esses algoritmos identificam *clusters* com base na densidade dos pontos de dados. Um *cluster* é formado quando a densidade na "vizinhança" excede um certo limite. Subcategorias incluem métodos baseados em regiões conectadas, ordenação de pontos para identificação de estrutura de *clusters* e *clusterização* baseada em funções de distribuição de densidade;
- Método Baseado em Tabela: Neste método, os objetos são alocados em células de uma tabela e agrupados com base na estrutura da tabela. Subcategorias incluem a Tabela de Informações Estatísticas e a *Clusterização* usando Transformação Wavelet;
- Método Baseado em Modelo: Para cada *cluster*, é criado um modelo hipotético para identificar quais dados se ajustam melhor ao modelo e localizam o *cluster*. Inclui a Maximização de Expectativa e o Modelo Conceitual;
- Método Baseado em Restrições: Utiliza restrições impostas pelo usuário ou pela aplicação para guiar a formação dos *clusters*;
- Método Baseado em Partições: Divide o conjunto de dados em k *clusters* distintos, onde cada objeto deve pertencer a exatamente um

cluster, e cada *cluster* deve conter pelo menos um objeto. Inclui subcategorias como K-medoids e K-means.

2.4.1 K-means

O método K-means agrupa um conjunto de n objetos em *k clusters* em um espaço *d*-dimensional. Cada *cluster* é representado por um centróide, e o objetivo é minimizar a distância euclidiana entre os objetos e seus respectivos centróides. O algoritmo busca minimizar a função objetiva conhecida como função da distorção quadrática, apresentada na Figura 4 a seguir:

Figura 4 - Fórmula da distorção quadrática

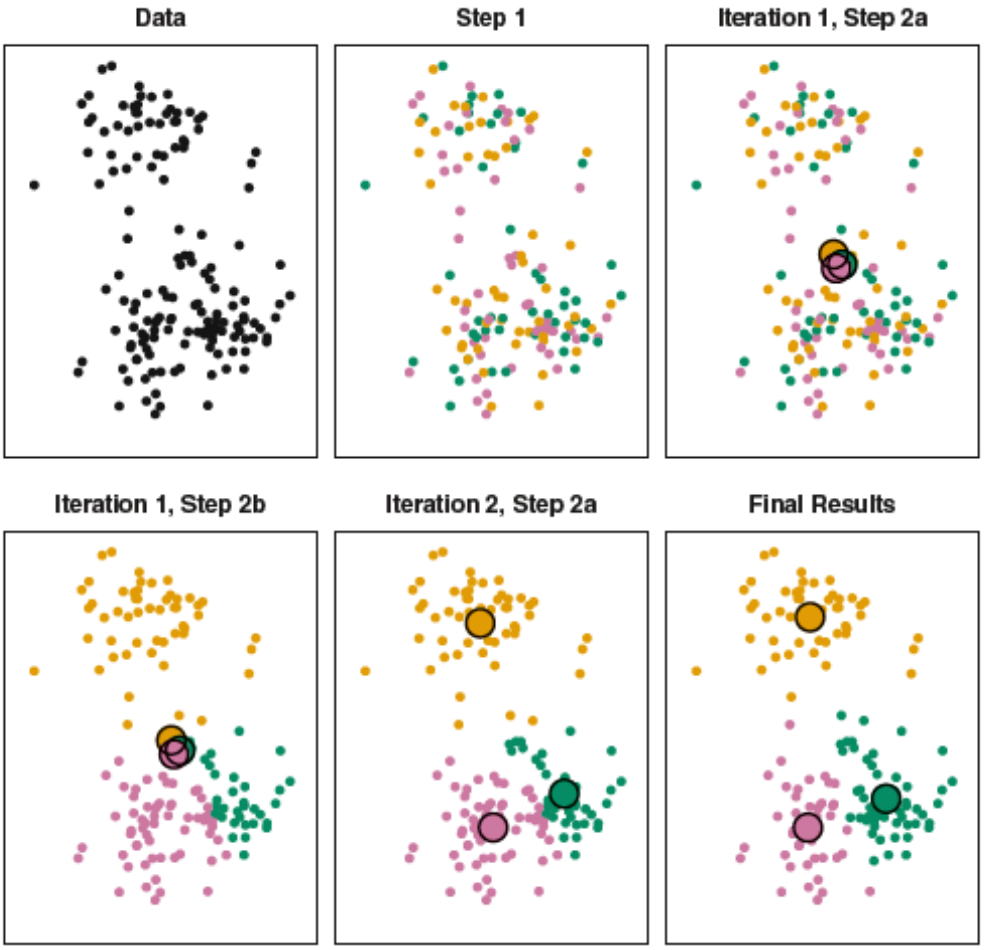
$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Fonte: (Pauletic; Prskalo; Bakaric, 2019, p.1398)

Onde “||x**i** - v**j**||” representa a distância euclidiana entre o objeto *x***i** e o centróide *v***j**.

Segue abaixo o exemplo visual da aplicação do algoritmo K-means em um conjunto de dados. A Figura 5 apresenta a clusterização com k=3.

Figura 5 - Representação da aplicação do algoritmo de clusterização K-means



Fonte: Gil (2024)

No primeiro retângulo da Figura 5, intitulado "*Data*", temos a visão geral da distribuição do dado, sem nenhum tipo de classificação.

No "*Step 1*", considerando que os dados já passaram por um processo de clusterização, onde cada dado foi atribuído um valor vetorial real, o algoritmo configurado com o K=3, reconhece a classificação dos vetores em seu respectivo *cluster* (rosa, verde ou amarelo).

No "*Iteration 1, Step 2a*", começa-se o processo de cálculo dos centróides dos clusters, primeiramente, este se apresenta localizado aleatoriamente.

No "*Iteration 1, Step 2b*", temos a separação dos dados, aglomerando-os conforme o *cluster* que estes participam, baseando-se na similaridade do valor dos vetores.

No "*Iteration 2, Step 2a*", temos a definição dos centróides de cada *cluster*.

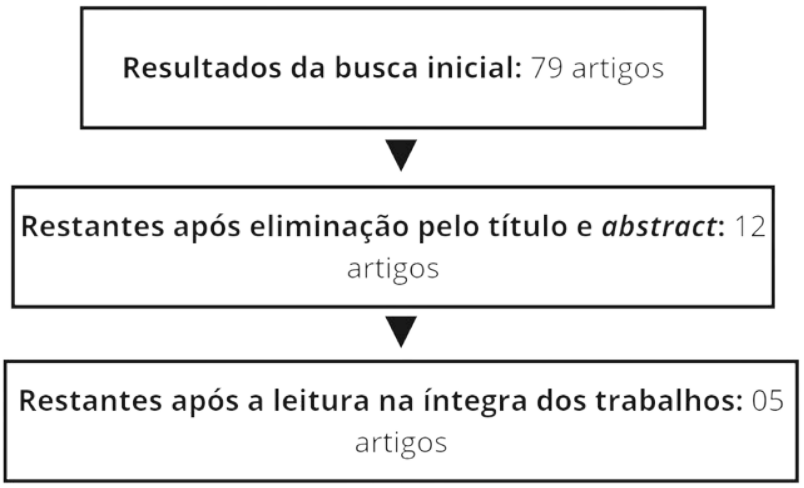
No "*Final Results*", temos a apresentação final da separação dos *clusters*.

2.5 TRABALHO CORRELATOS

Por meio de uma pesquisa abrangente na literatura científica, foram identificadas e selecionadas publicações acadêmicas que abordam a implementação de algoritmos para tratamento ou geração de ideias, com foco em algoritmos de aprendizado de máquina (LLM ou outros) na gestão de ideias. A pesquisa foi conduzida nas bases de dados SCOPUS® e IEEE®, considerando trabalhos publicados na última década em inglês, utilizando a seguinte string de busca: *("Idea Visualization" OR "Idea Management" OR "Idea Generation" OR "Idea Mining") AND ("Language Models" OR "Pretrained Models" OR "Natural Language Processing" OR "NLP Models" OR "Text Embeddings" OR "Machine Learning")*.

A Figura 6 ilustra o processo de seleção dos trabalhos mais alinhados ao tema desta pesquisa.

Figura 6 - Passos executados na identificação dos trabalhos correlatos



Fonte: Autora (2024).

Inicialmente, foram encontrados 79 artigos relevantes, dos quais 26 apresentavam títulos e resumos relacionados à temática proposta. Após uma leitura detalhada dos textos restantes, foram selecionados 5 trabalhos que demonstraram maior similaridade com a pesquisa em questão, permitindo a continuidade da discussão. O comparativo entre o presente estudo e os trabalhos selecionados é apresentado no Quadro 1, que inclui um resumo de cada artigo.

O artigo elaborado por Cui *et al.* (2024) traz a comparação dos resultados obtidos através da aplicação de três métodos diferentes de *data mining* em um

mesmo conjunto de dados, a fim de compreender quais destes teriam um melhor e maior impacto na extração de *insights*. O conjunto de dados em questão é um compilado de *reviews* sobre o produto “aspirador de pó Dyson” extraído da rede social chinesa “*Little Red Book*”, onde neste, os autores buscaram encontrar os *reviews* mais significativos, a fim de obter as sugestões - *ideation* - mais relevantes para o processo de inovação do produto. Os algoritmos utilizados foram *isolation forest*, análise de *cluster* baseado em densidade, e método *autoencoder*. Os pesquisadores analisaram que os métodos de *machine learning isolation forest* e análise de agrupamento baseado em densidade são muito sensíveis ao tamanho das sugestões (pois estas variam de 10 à 2150 palavras). Já o método *deep Learning Autoencode*r mostrou estabilidade e capacidade de detectar sugestões únicas do *dataset* como um todo.

Já os autores Cheddak *et al.* (2024), explicitaram o desenvolvimento de um sistema que aborda o processo de *ideation* que acontece durante sessões de *brainstorm,* extraindo de uma sessão com o tótipo “fracasso acadêmico”, ideias através de *Idea Mining.* O dataset foi construído a partir da leitura dos post-its utilizados na sessão com OCR, onde foi aplicado o processo de *data augmentation* com o modelo de linguagem GPT-3.5 para garantir a robustez do texto. Após limpo e normalizado, foi aplicado o algoritmo BERTopic, que engloba vetorização (AraBERT-v02), redução de dimensionalidade (UMAP), clusterização (HDBSCAN), extração de tópicos (c-TF-IDF) e representação visual com palavras-chave. Os resultados obtidos através da modelagem proposta se mostraram atraentes, na visão dos autores, pois se tornam uma ferramenta para os gerenciadores destas dinâmicas, classificando de maneira automática e apontando os termos relevantes, aumentando assim, a eficiência do processo.

Mahdi *et al.* (2022), trouxeram a perspectiva da aplicação de um modelo de camadas para o processamento de sugestões. Este modelo se baseia em um pré-processamento dos dados para remoção de ruídos, aplicação de BERT e K-Means, construção de um classificador de sugestões treinado pelos autores, e por fim, aplicação de similaridade de cosseno. O estudo usou como fonte de dados um *dataset* de *reviews* de comida da *Amazon Food*. Aplicou-se o classificador para remover do dataset todas as linhas que não se tratavam de sugestões, e para o estudo, considerou-se apenas as primeiras 2000 sugestões com accuracy superior à 0.9. Por fim, estas foram utilizadas no algoritmo de similaridade (cálculo de

cosseno). Os resultados encontrados foram positivos na visão dos autores, pois estes apontaram que a junção de tais tecnologias melhoraram o processo de resgate e classificação das sugestões, bem como as ideias resgatadas traziam *insights* interessantes.

Hoornaert *et al*. (2017) apresentaram em seu trabalho a modelagem de um sistema de ranqueamento de ideias e sugestões mais passíveis de serem implementadas, provindo de fontes públicas diversas, como *feedbacks* diretos, *suggestion box*, etc, para apoiar gestores em seus negócios, com o viés “3 C’s” - conteúdo, o contribuidor que propôs e o feedback público sobre a ideia. Através da implementação do método de extração de informação *latent semantic indexing,* os autores resgataram um *dataset* de sugestões e ideias relacionadas ao *Mendeley*. A partir disto, aplicaram testes com métodos lineares como o *linear discriminant analysis* e o *regularized logistic regression*; e métodos não lineares como os machine learning *stochastic adaptive boosting* e *random forest*. Os resultados obtidos foram classificados e analisados em “*Real-Time Data*” e “*Time-Delayed Data*”, isso para compreender se o algoritmo era capaz de tomar boas decisões com dados em tempo real, ou se uma análise temporal dos mesmos era necessária. Com isso, os autores perceberam que os resultados observados no grupo “*Time-Delayed Data*” eram mais valiosos e possuíam uma maior acurácia, se comparado com “*Real-Time Data*”.

Já o artigo redigido por Lee *et al*. (2017) propõe um método de análise de ideias, que busca ultrapassar a dificuldade do *overload* de dados em um cenário de um ambiente de inovação aberto. Para isso, estes escolheram como base de dados o site *MyStarbucksIdea.com*, e avaliaram a probabilidade de adoção da ideia sugerida, através de uma dupla análise. De modo separado, foi utilizado em um primeiro *dataset* a aplicação do algoritmo TF-IDF, atribuindo às palavras pesos, e posterior a isso, o processo de tokenização, normalização e reconhecimento. No segundo *dataset,* foi levado em consideração aspectos como números de ideias que aquele mesmo usuário submeteu, número de votos recebidos, etc. Ainda, os autores aplicaram a análise de sentimento *SentiWordNet*, pois julgaram que este aspecto influenciou na relevância da ideia. Por fim, estes combinaram os *outputs* dos dois *datasets*. Como resultado, os autores perceberam que a melhor saída obtida pelo modelo proposto foi quando estes rodam testes em cima dos *datasets* combinados.

Quadro 1 - Comparação entre os trabalhos correlatos e o presente trabalho

Autores	Ano	Algoritmo utilizado	Conjunto de dados	Resultados
Cui *et al.* (2024)	2024	*Isolation forest*, análise de *cluster* baseado em densidade, e método autoencoder	Compilado de reviews sobre o produto "aspirador de pó Dyson" extraído da rede social chinesa "Little Red Book"	O método *deep Learning Autoencode*r foi quem apresentou estabilidade e capacidade de detectar sugestões únicas do dataset como um todo.
Cheddak et al.	2024	BERTopic (AraBERT-v02, UMAP, HDBSCAN, c-TF-IDF)	Construído através da leitura através de OCR de *post-its*, após uma sessão de *brainstorm.*	Os resultados obtidos orbitam o aumento da eficiência do processo de brainstorm, pela agilidade que o sistema traz.
Mahdi *et al*	2022	Módulo de transformação de sentenças baseado em BERT - 'paraphrase-distilroberta-base-v2 [com *built-in* K-Means]	Dataset de *Reviews* da Amazon *Food*	O método de vetorização BERT se mostrou eficiente quando combinado com as técnicas de text mining escolhidas, apresentando uma boa *accuracy*.
Hoornaert *et al.*	2017	*Linear discriminant analysis, regularized logistic regression*; *machine learning stochastic adaptive boosting* e *random forest.*	Dataset de sugestões e ideias relacionadas ao Mendeley	A aplicação dos algoritmos no set "Time-Delayed Data" obteve maior acurácia, se comparado ao "Real-Time Data".
Siangliulue *et al*	2017	TF-IDF e *SentiWordNet*	Comunidade aberta de sugestões "*MyStarbucksIdea.com*"	A saída que apresentou melhor *output* foi através da combinação dos datasets.

Fonte: Autora (2024).

3 METODOLOGIA

Este capítulo tem como objetivo fornecer uma descrição dos métodos e técnicas empregados na condução da pesquisa.

3.1 CARACTERIZAÇÃO DA PESQUISA

Os autores Gerhardt e Silveira (2009) apontam que uma pesquisa inicia através do estabelecimento de uma problemática ou dúvida, sendo necessário um método ou caminho a ser trilhado para a obtenção de algum resultado, gerando então conhecimento. Assim, este trabalho seguirá uma metodologia aplicada e exploratória, com intuito de extrair de um conjunto de dados, *insights*.

A pesquisa aplicada, busca gerar conhecimento através da resolução de alguma problemática inicial, implicando no aprimoramento científico e inovação deste tema (VIEIRA; LEITE; KUHN, 2023). Já a pesquisa exploratória é empregada frequentemente em temas pouco explorados, visando trazer um apanhado geral da mesma, permitindo aos pesquisadores uma proximidade com as peculiaridades do tema (GIL 2008).

Nesse contexto, este trabalho permeia as duas esferas citadas, através da aplicação de algoritmos LLM em uma base de dados de ideias, buscando assim explorar os *insights* e tendências que possam revelar oportunidades de mercado e inovação nos diversos setores econômicos.

Ainda, foi conduzida uma pesquisa bibliográfica sobre o tema proposto neste trabalho, com intuito de fundamentar teoricamente o mesmo. Segundo Lima e Mioto (2007) a pesquisa bibliográfica fornece informações e elementos que assistem o processo de análise dos dados obtidos e seus resultados.

A próxima seção detalha a metodologia utilizada nesta pesquisa.

3.2 METODOLOGIA *DESIGN SCIENCE RESEARCH METHODOLOGY*

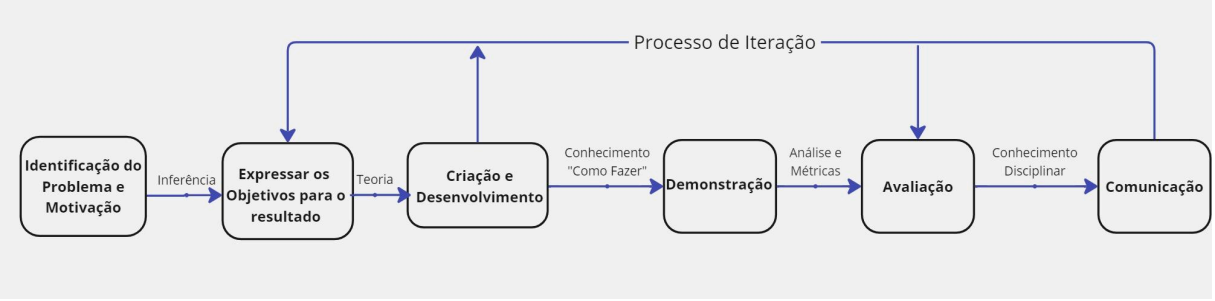
A metodologia DSRM (*Design Science Research Methodology*), tem como seu objetivo aprimorar sistemas existentes, resolver problemas e criar novos artefatos (Peffers et al., 2007).

Os autores Abdurrahman e Mulyana (2020) apontam que por mais que esta metodologia siga um fluxo de execução pré definido, isso não impede de adaptação por parte da mesma, a depender do objetivo e resultado final esperado. A seguir, uma breve descrição de cada passo será apresentada:

- Passo 1 - Identificação do problema e motivação: atém-se na definição e descrição do problema de pesquisa, bem como na justificativa da solução proposta e seu valor;
- Passo 2 - Expressar os objetivos para o resultado: explicita a raciocínio final das ideias obtidas a partir da problemática definida anteriormente, assim como as ações cabíveis a pesquisa;
- Passo 3 - Criação e desenvolvimento: Diz respeito a criação de um artefato, podendo ser este um framework, um método, uma arquitetura etc, que venha a determinar quais as funcionalidades que o artefato real terá;
- Passo 4 - Demonstração: apresenta o uso do artefato como resolução do problema de maneira inteira ou parcial, dentro de um cenário de teste;
- Passo 5 - Avaliação: trata-se da observação e avaliação do uso do artefato para solucionar o problema no cenário proposto;
- Passo 6 - Comunicação: este passo é responsável por divulgar o problema, a modelagem feita e os resultados encontrados com o artefato, bem como a sua eficácia.

A Figura 7, apresenta o fluxograma de execução nominal da DSRM.

Figura 7 - fluxo de execução nominal Design Science Research Methodology



Fonte: Adaptado de Peffers *et al.*(2007, p.54)

3.3 DESENVOLVIMENTO DA PESQUISA

A fim de dar seguimento na pesquisa, a implementação da DSRM foi esquematizada no Quadro 2 a seguir, com as devidas adaptações.

Quadro 2 - Metodologia DSRM

Identificação do problema e motivação	Explorar os dados coletados de sites de geração de ideias, identificando a necessidade de ferramentas que facilitem a análise automatizada de grandes volumes de dados textuais, com foco na identificação de padrões e tendências inovadoras.
Definição dos requisitos	Estabelecer os critérios para a análise exploratória das ideias coletadas, incluindo a identificação de grupos temáticos, padrões semânticos e critérios de avaliação para a eficácia da clusterização, a fim de compreender o panorama tecnológico e inovador abordado pelos sites analisados.
Projeto e desenvolvimento	Desenvolver e implementar um processo de clusterização, utilizando algoritmos de agrupamento, como K-Means, integrados com Modelos de Linguagem de Larga Escala (LLMs) para identificar e organizar ideias semanticamente e contextualmente semelhantes, facilitando a análise de grandes volumes de dados textuais.
Demonstração	Aplicar o modelo de LLM e o algoritmo K-Means para processar e analisar os dados coletados, gerando visualizações dinâmicas que representem os agrupamentos identificados e os principais *insights* extraídos das ideias coletadas.
Avaliação	Avaliar os *clusters* gerados, verificando a coerência e relevância dos agrupamentos identificados.
Comunicação	Apresentar os resultados obtidos em relatórios e visualizações que destaquem os *insights*, padrões e tendências identificadas. Discutir o impacto desses achados para o campo da inovação, sugerindo potenciais aplicações práticas dos *clusters* gerados para otimizar a gestão de ideias nas organizações.

Fonte: Autora (2024).

3.4.1 População e Amostra

Para viabilizar o estudo proposto, optou-se por utilizar uma fonte de dados composta por ideias extraídas de sites de geração de ideias. Como população deste estudo, foi selecionado um site especializado na divulgação de ideias de negócios, que funciona como um repositório online abrangendo diversas categorias de ideias.

Com o objetivo de realizar uma análise completa das ideias disponíveis, decidiu-se trabalhar com todas as categorias apresentadas no site, considerando que cada uma delas desempenha um papel importante na identificação de tendências e mudanças nos setores econômicos abordados. Dessa forma, a amostra utilizada para análise compreendeu um total de 9.698 ideias, distribuídas em 24 categorias específicas e uma categoria adicional denominada "*Uncategorized*", contendo nove ideias não categorizadas no dataset.

O período de coleta das ideias abrangeu de maio de 2004 a abril de 2024, oferecendo uma perspectiva de duas décadas de evolução e inovação em diversos setores. A análise buscou, por meio de visualizações, explicitar os componentes do *dataset*, além de realizar buscas semânticas para identificar ideias relacionadas a temas semelhantes. Adicionalmente, aplicaram-se buscas vetoriais nas categorias específicas, com o intuito de encontrar as ideias mais semelhantes dentro de cada categoria.

3.4.2 Coleta de Dados e Procedimentos

O site de geração de ideias utilizado não disponibilizava uma API ou um conjunto de dados pré-definido para o acesso às suas informações. Portanto, a coleta de dados foi realizada por meio do método de *web scraping*, que consiste na extração automatizada de dados de páginas *web*. Neste estudo, foram coletados apenas os resumos das postagens disponíveis no site, devido ao seu formato conciso, que proporciona uma visão geral de cada ideia. O processo de coleta de dados seguiu as seguintes etapas:

- Navegação: Inicialmente, realizou-se a navegação manual pelo site de geração de ideias para identificar as páginas a serem extraídas;

- Utilização de Biblioteca Específica: Para a extração dos dados das páginas HTML, utilizou-se a biblioteca *BeautifulSoup*, que facilita a navegação e coleta de informações estruturadas;
- Seleção dos Dados: Foram definidos os campos essenciais para compor o *dataset*, incluindo título da postagem, resumo, data de publicação, entre outros;
- Configuração do Paginador: O *script* foi configurado para percorrer múltiplas páginas do site, garantindo a coleta de todas as postagens disponíveis;
- Início do processo de *Web Scraping*: O processo de extração dos dados foi iniciado com a leitura de *tags* específicas do HTML, que continham as informações relevantes;
- *Download* do Dataset: Após a conclusão do processo de coleta, os dados extraídos foram estruturados e armazenados em um arquivo CSV para posterior análise.

3.4.3 Análise de dados

A fase de análise de dados é fundamental para a extração de *insights* e tendências a partir do dataset. Nesta pesquisa, foi adotada uma abordagem multifacetada, utilizando análise de agrupamento, processamento de linguagem natural e diversas técnicas de visualização de dados para agrupar, interpretar e explorar as ideias de negócio disponíveis nas várias categorias do site.

Os principais procedimentos adotados na análise de dados foram:

- Geração de *Embedding*: Utilizou-se o modelo BERT para vetorização das ideias. Esse método permite a transformação das ideias em representações vetoriais, capturando o significado semântico das mesmas;
- Agrupamento com o algoritmo K-Means: As ideias foram agrupadas com base em sua similaridade utilizando o algoritmo K-Means. O número de *clusters* (*k*) foi definido por meio do método do cotovelo, que identifica o ponto ideal onde a adição de novos *clusters* não melhora significativamente a variabilidade explicada;
- Redução de Dimensionalidade com PCA: Para simplificar a análise inicial e facilitar a visualização dos *clusters*, foram aplicadas as técnicas de Análise de

Componentes Principais (PCA). Essa técnica reduz a dimensionalidade dos dados vetorizados, permitindo a observação de padrões de agrupamento de forma mais intuitiva;

- Busca Semântica: Modelos de linguagem de larga escala (LLMs) foram utilizados para realizar buscas semânticas dentro dos *clusters*, com o objetivo de encontrar ideias que se relacionassem a palavras-chave específicas, facilitando a identificação de temas ou padrões recorrentes;
- Visualização de Dados: Foram criados diferentes gráficos e representações visuais para ilustrar a composição e as características de cada *cluster*. Essas visualizações auxiliaram na interpretação tanto dos aspectos qualitativos quanto quantitativos dos dados agrupados, oferecendo uma visão mais detalhada das tendências identificadas.

3.4.4 Limitações da pesquisa

Este estudo enfrentou limitações devido ao teor das ideias, onde entende-se que o site selecionado não compreende a totalidade de temas e ideias de negócios.

3.4.5 Resultados esperados

Ao final desta pesquisa, espera-se alcançar os seguintes resultados:

- Identificar *insights* e tendências: O objetivo é identificar, nas diversas categorias de ideias, as principais tendências, os temas mais abordados e as categorias com maior destaque, fornecendo *insights* significativos que possam apoiar decisões estratégicas relacionadas à inovação;
- Utilizar buscas semânticas com modelos de linguagem de larga escala (LLMs): para melhorar a precisão na recuperação de ideias relacionadas a temas específicos;
- Aprimoramento da Gestão de Ideias: Com base nos *insights* identificados, espera-se encontrar valor significativo que impacte diretamente a gestão de ideias, melhorando a forma como estas são obtidas, processadas e tratadas.

O objetivo é contribuir para a gestão da inovação, facilitando a transformação de ideias em ações concretas.

Além desses resultados, este estudo também tem o potencial de contribuir academicamente, ao proporcionar um entendimento mais aprofundado dos padrões semânticos presentes nas ideias analisadas, ao utilizar métodos de processamento de linguagem natural e ao possibilitar a descoberta de padrões e *insights*.

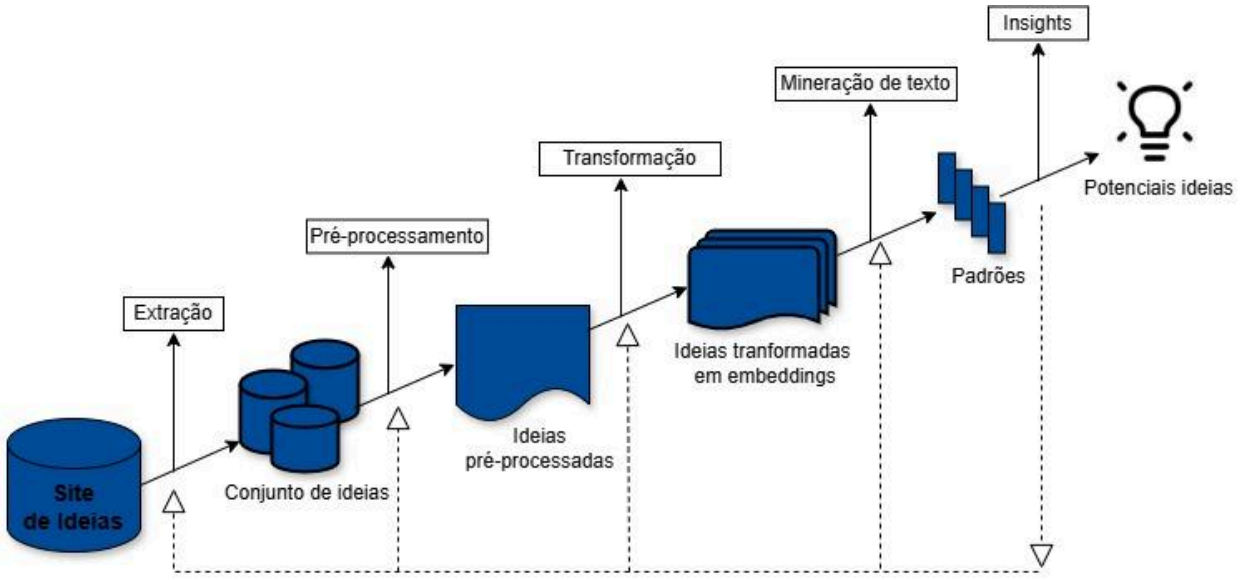
4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

O presente capítulo tem como objetivo explicitar o método utilizado neste trabalho, detalhar os principais resultados obtidos e discutir os *insights* gerados. A análise compreendeu um total aproximado de 10.000 ideias (dez mil), classificadas em 25 categorias distintas ao longo de um período de duas décadas. A abordagem metodológica adotada englobou a aplicação de técnicas de extração, limpeza, vetorização e agrupamento de dados (*embedding e clustering*), além de métodos para redução de dimensionalidade e visualização dos dados processados.

4.1 MÉTODO PROPOSTO

A Figura 8 apresenta o fluxo esquematizado do método implementado, ilustrando as etapas principais do processo. Em seguida, cada fase é descrita em detalhes, conforme os passos a seguir:

Figura 8 - Método proposto



Fonte: Adatpado de Fayyad; Piatetsy-shapiro; Smith (1996).

4.1.1 Extração

Após a escolha do site a ser utilizado, o primeiro passo consistiu na extração dos dados da plataforma. Utilizou-se a técnica de *web scraping* para resgatar informações como categoria, título, conteúdo, *link* de acesso e data de publicação. Os dados coletados foram armazenados em um arquivo no formato *CSV*, permitindo a sua manipulação nas etapas subsequentes.

4.1.2 Pré-processamento

Nesta fase, o foco foi aprimorar a qualidade dos dados extraídos, otimizando o processo de análise posterior. Primeiramente, os dados foram carregados a partir do *dataset* salvo na etapa anterior, utilizando a biblioteca *Pandas*, conforme ilustrado na Figura 9.

Figura 9 - *Raw dataset*

	post-categories	entry-title	entry-title href	entry-content	entry-date
0	Business Trends	The Ultimate Guide to Recession-Proof Business...	https://www.coolbusinessideas.com/archives/the...	For entrepreneurs, navigating the fluctuating ...	2024-04-17
1	Society & Environment	Magorium Recycled Roads	https://www.coolbusinessideas.com/archives/mag...	A company named Magorium has developed a way t...	2024-03-17
2	Society & Environment	WaveRoller Wave Energy Converter	https://www.coolbusinessideas.com/archives/wav...	A Finnish company, AW-Energy, has developed a ...	2024-03-16
3	Travel & Transport	Robotic, on-demand system for charging electri...	https://www.coolbusinessideas.com/archives/rob...	Kolbev's system uses robots to deliver chargin...	2024-03-15
4	Online & Social Networks	Sick of Google? 5 Search Engine Alternatives t...	https://www.coolbusinessideas.com/archives/sic...	While Google remains dominant in the search en...	2024-01-29

Fonte: Autora (2024).

O processo de limpeza e normalização dos dados incluiu:

1. Normalização textual: As colunas *"entry-title"* e *"entry-content"*, que contêm os títulos e conteúdos das ideias, foram convertidas para letras minúsculas e tiveram os caracteres especiais removidos, garantindo padronização e minimizando erros de processamento;
2. Concatenação de colunas: Criou-se uma nova coluna chamada "texto", resultante da união das colunas *"entry-title"* e *"entry-content"*, oferecendo uma visão consolidada de cada ideia para fins de análise.

3. Formatação de datas: A coluna *“entry-date”* foi ajustada para o padrão de data AAAA-MM-DD, assegurando consistência em todo o dataset.
4. Remoção de duplicatas: Garantiu-se a unicidade das ideias removendo entradas duplicadas com base na coluna “texto”.
5. Após essas etapas, obteve-se um dataset limpo, estruturado e pronto para ser transformado nas fases subsequentes.

4.1.3 Transformação

Com o dataset preparado, iniciou-se a etapa de transformação, que envolveu o processo de vetorização (*embedding*). Utilizou-se o método BERT (*Bidirectional Encoder Representations from Transformers*) para gerar representações vetoriais dos textos. Essa abordagem considera o contexto das palavras, permitindo que uma mesma palavra assuma valores vetoriais distintos conforme sua posição e significado na frase.

A vetorização foi implementada com a biblioteca *SentenceTransformer*, e os dados transformados foram armazenados no arquivo *CSV* denominado *“embedding_train.csv”*, ilustrado na Figura 10.

Figura 10 - *Dataset* de *embeddings*

```
df_embedding.to_csv("embedding_train.csv",index = False)
```

Batches: 100% 303/303 [07:59<00:00, 1.32s/it]

```
df_embedding
```

	0	1	2	3	4	5	6	7	8	9	...	374	375	376	377	3
0	-0.070606	-0.015561	-0.037481	-0.055326	0.061289	0.062277	-0.067507	-0.008687	-0.029147	-0.019469	...	0.048913	-0.058186	-0.000411	0.040956	0.0461
1	-0.085534	0.011283	-0.048736	0.031898	0.033984	-0.106366	0.020758	0.030942	-0.100295	-0.032542	...	0.040809	0.052848	0.026639	0.007909	-0.0426
2	-0.113797	-0.013650	-0.043195	0.001315	0.011126	-0.031783	-0.042622	0.040713	-0.028479	-0.006381	...	0.011241	-0.032566	-0.017990	-0.014090	-0.0624
3	-0.129228	-0.057678	-0.118894	-0.010925	0.066408	-0.019691	-0.000059	0.002998	-0.009145	0.014139	...	0.171603	-0.009647	-0.001794	-0.008863	-0.0917
4	-0.097084	-0.096790	0.044814	-0.015524	-0.033409	-0.045626	-0.030829	0.000948	-0.004115	-0.104227	...	0.037315	-0.002481	0.020602	-0.016234	-0.0652
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9668	-0.012795	-0.062104	-0.040038	-0.071045	0.007844	0.001073	0.046075	0.054489	-0.024309	0.018554	...	0.071778	-0.035047	0.030764	-0.019172	-0.0361
9669	-0.055820	-0.011428	-0.032632	-0.046061	0.026333	-0.080030	-0.022931	0.001045	-0.041358	-0.019058	...	-0.026613	0.037002	0.086239	-0.007618	0.0115
9670	-0.026105	-0.006752	0.022419	-0.042809	-0.100342	-0.066478	0.060524	0.056268	-0.003536	-0.038902	...	0.011550	-0.038062	0.029297	0.038795	0.0308
9671	-0.060199	-0.036013	-0.021335	-0.024581	-0.060735	-0.055565	-0.031110	-0.068442	-0.000740	-0.003876	...	0.047933	-0.089669	0.000415	0.017368	0.0225
9672	-0.010792	-0.058885	-0.022877	-0.071968	0.042071	-0.056838	0.030463	-0.055722	-0.041858	-0.010674	...	0.021622	-0.070713	-0.054998	0.056931	0.0186

9673 rows × 384 columns

Fonte: Autora (2024).

4.1.4 Mineração de texto

Na etapa de mineração de texto, foi aplicada a técnica de redução de dimensionalidade utilizando o método PCA (*Principal Component Analysis*). Essa abordagem estatística permite reduzir a dimensionalidade vetorial, mantendo os aspectos mais relevantes para a análise.

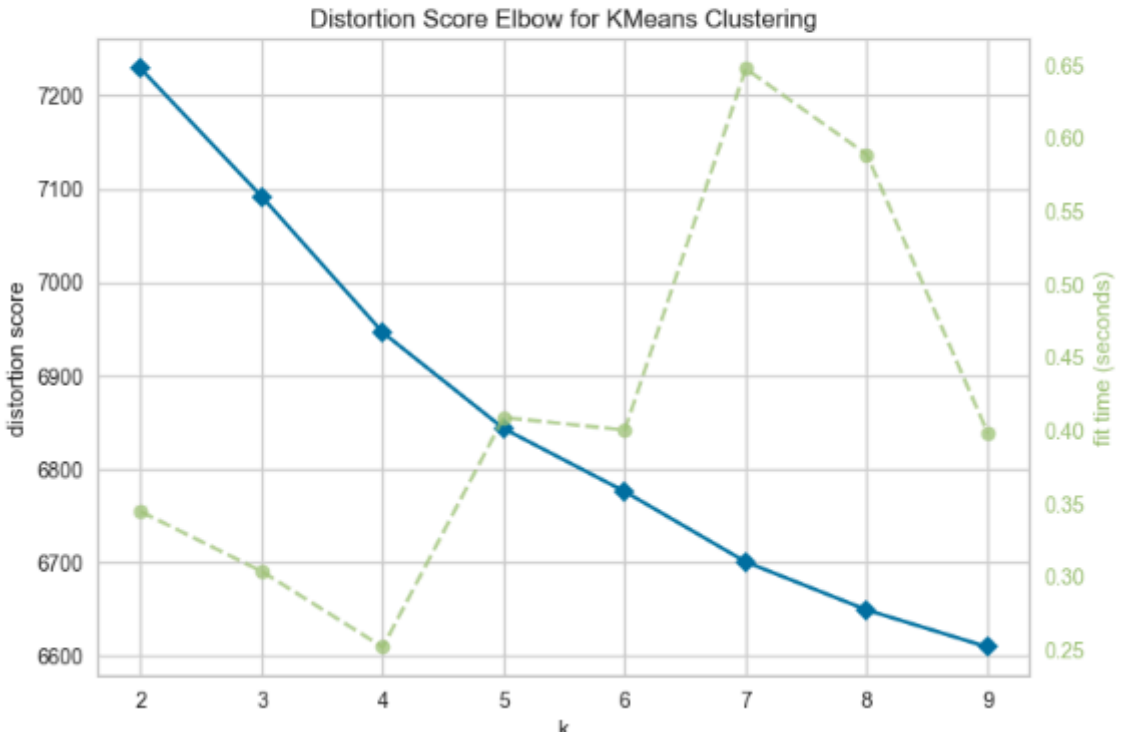
Dois passos principais foram realizados:

1. Remoção de *outliers*: Anomalias ou casos atípicos foram excluídos para evitar distorções nos resultados.
2. Definição do número ideal de *clusters*: Por meio do método do cotovelo (*elbow method*), calculou-se a quantidade ideal de *clusters*, conforme mostrado na Figura 11.

Figura 11 - Método cotovelo

```
km = KMeans(init="k-means++", random_state=0, n_init="auto")
visualizer = KElbowVisualizer(km, k=(2,10), locate_elbow=False)

visualizer.fit(df_embedding_with_out)
visualizer.show()
```



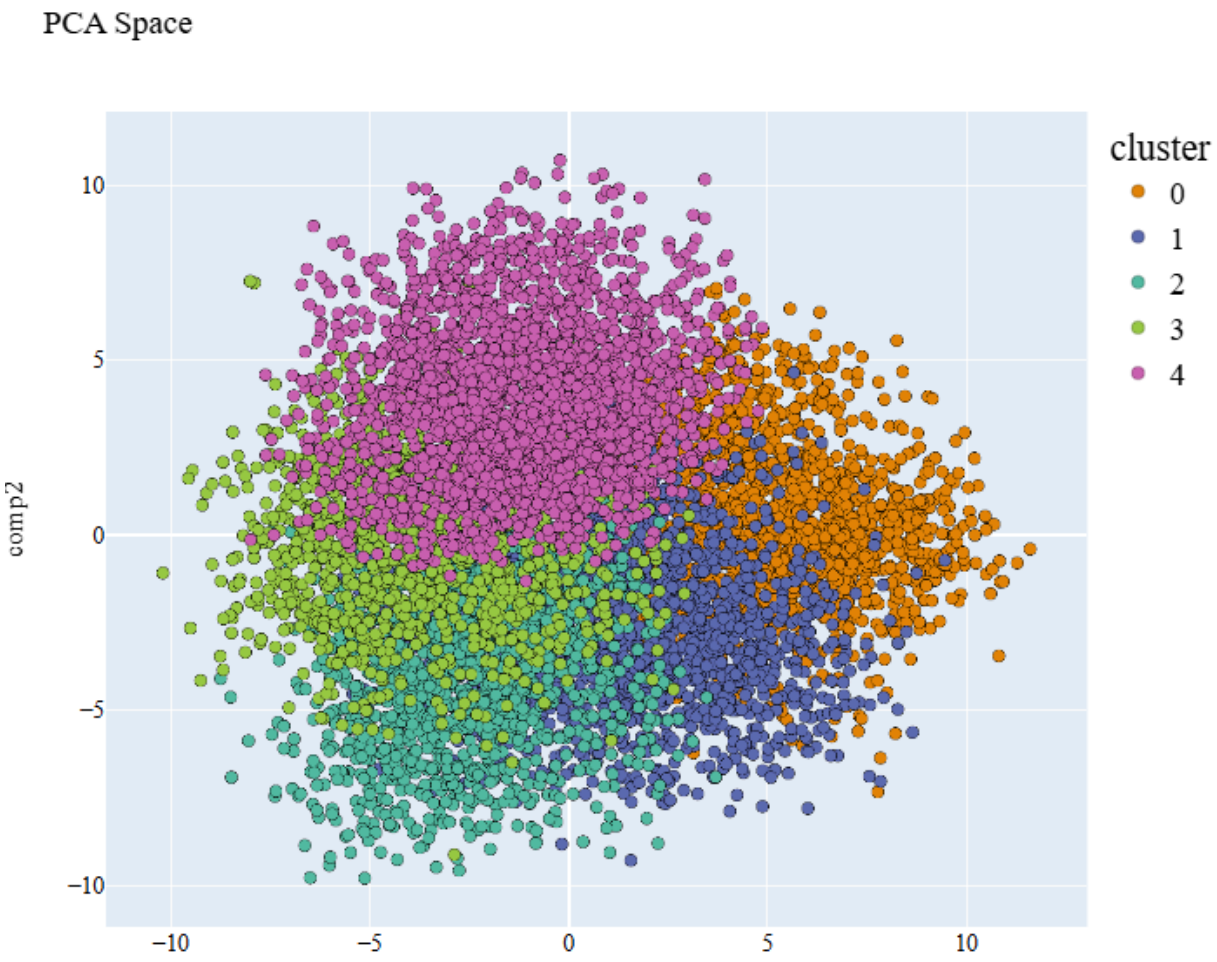
Fonte: Autora (2024).

O método do cotovelo tem como função auxiliar na avaliação da variação dos dados do conjunto estudado em relação ao número de agrupamentos possíveis. No contexto deste trabalho, foi decidido utilizar o número de *clusters* igual a cinco (*K=5*),

com base na observação de que essa escolha oferece uma representação mais equilibrada e harmoniosa das ideias dentro de seus respectivos grupos.

Posteriormente, o algoritmo *K-Means* foi aplicado para agrupar os vetores com *K=5*, conforme determinado pelo método do cotovelo. A partir disso, geraram-se plotagens utilizando as técnicas de PCA (*Principal Component Analysis)*, em duas e três dimensões, visando facilitar a interpretação visual dos dados contidos no dataset. A Figura 12 ilustra a plotagem 2D obtida com o PCA.

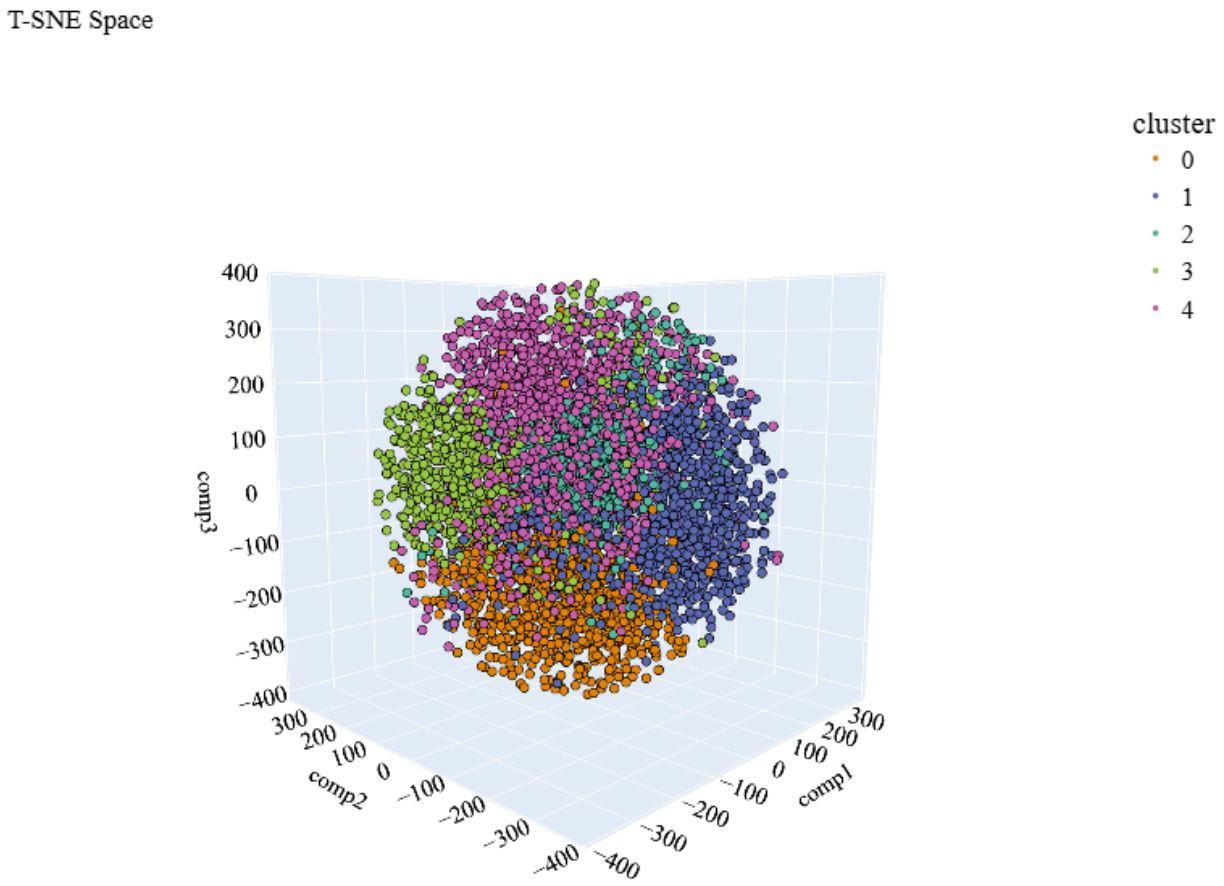
Figura 12 - PCA 2D



Fonte: Autora (2024).

Em complemento, a Figura 13 apresenta a plotagem realizada com o t-SNE (*t-distributed Stochastic Neighbor Embedding*). Diferentemente do PCA, que prioriza a redução da dimensionalidade destacando os principais componentes, o t-SNE reduz a dimensionalidade preservando as similaridades locais entre pares de dados no conjunto, o que contribui para uma análise mais detalhada das proximidades conceituais.

Figura 13 - *T-SNE*



Fonte: Autora (2024).

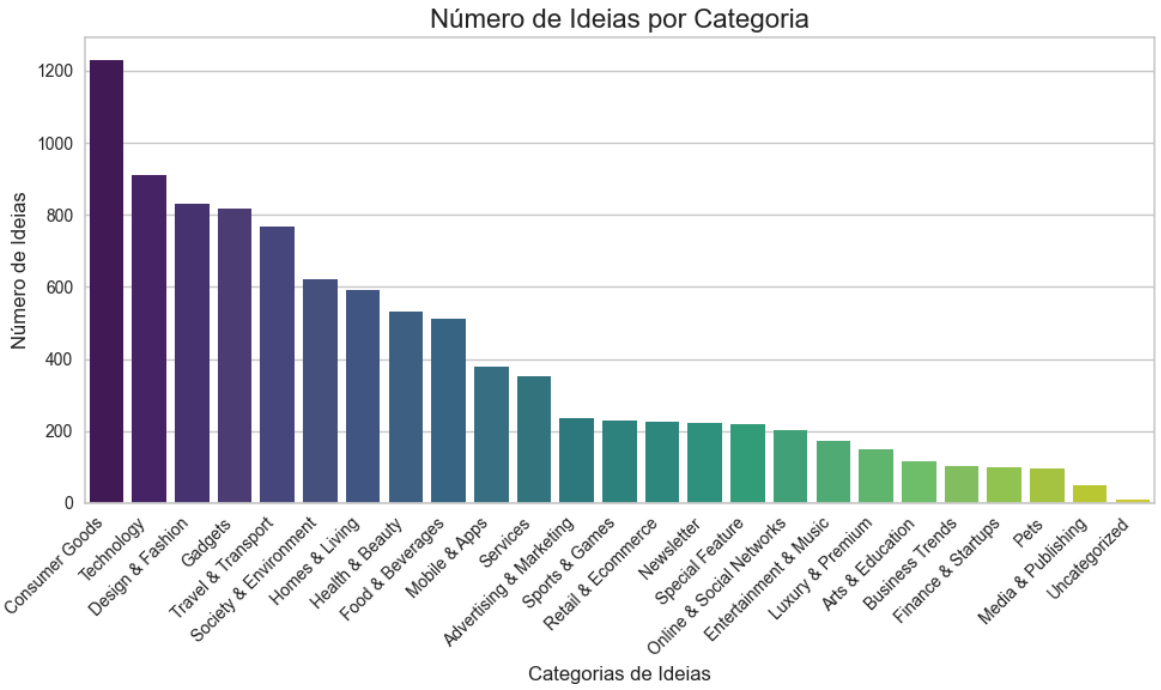
4.2 RESULTADOS OBTIDOS

Os resultados obtidos serão apresentados em dois momentos. O primeiro, apresentará os dados volumétricos, qualitativos e análises temporais realizadas. O segundo trará as análises de busca semântica.

4.2.1 Visualização de dados - categorias

As visualizações subsequentes têm como objetivo oferecer uma visão geral do dataset. A Figura 14 mostra as categorias presentes no conjunto de dados, organizadas em ordem decrescente conforme a quantidade de ideias vinculadas a cada uma delas.

Figura 14 - Número de Ideias por Categoria



Fonte: Autora (2024).

As categorias com maior número de ideias são *“Consumer Goods”, “Technology”* e *“Design & Fashion”*, cada uma contendo mais de 800 ideias. Para aprofundar a análise, foram geradas nuvens de palavras, destacando os principais termos associados a cada categoria. Essas nuvens revelam tendências e áreas de interesse específicas.

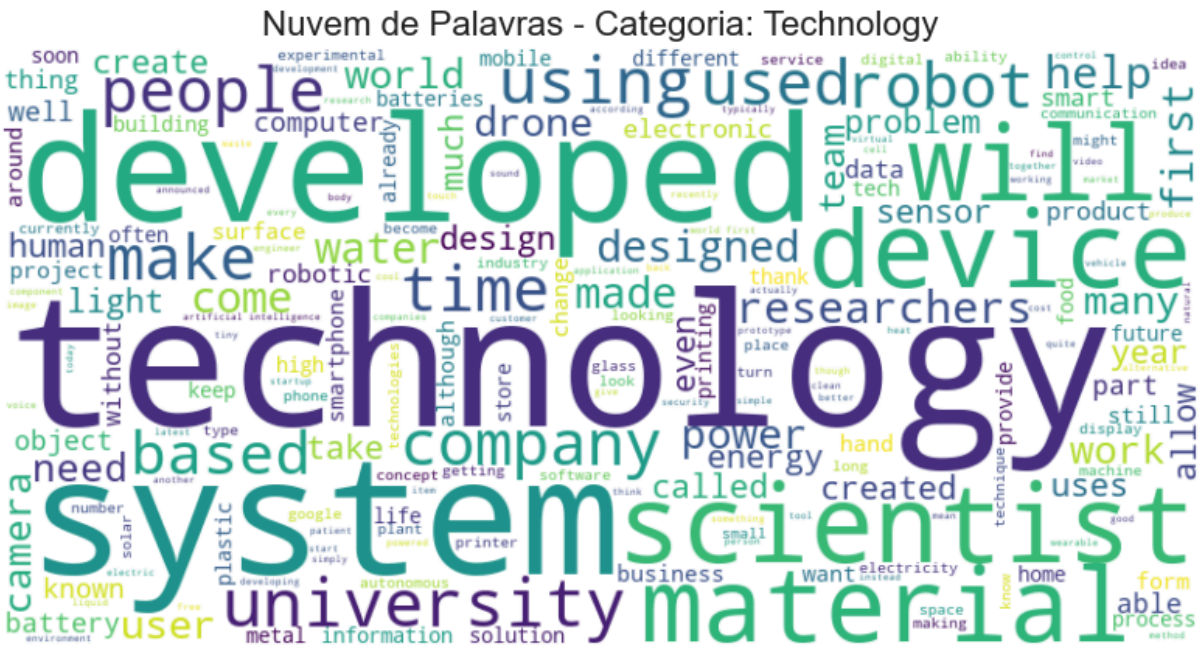
Na categoria *“Consumer Goods”*, a Figura 15 destaca termos como *“designed”, “product”* e *“time”*, entre outros, sugerindo que as ideias associadas a essa categoria estão voltadas para produtos, dispositivos e sistemas inovadores, frequentemente focados em atender às necessidades humanas de forma inteligente e eficiente. Esses termos também evidenciam a importância do *design* pensado para o usuário, da gestão do tempo e de soluções que promovam praticidade e conectividade, alinhando-se às demandas de um mercado cada vez mais orientado à experiência do consumidor.

Figura 15 - Nuvem de Palavra referente à categoria “Consumer Goods”

Fonte: Autora (2024).

A categoria *“Technology”* é caracterizada por palavras-chave como *“technology”, “developed”, “system”, “device”, “university”* e “*scientist”*, entre outros termos que destacam o papel da tecnologia em diversas esferas. Esses termos apontam para o desenvolvimento e aplicação de sistemas e dispositivos inovadores, tanto no contexto acadêmico, com pesquisas conduzidas por universidades e cientistas, quanto no cotidiano das pessoas, através de ferramentas tecnológicas que facilitam a vida moderna. Além disso, a presença desses termos sugere uma interseção entre inovação, pesquisa e acessibilidade, indicando a crescente influência da tecnologia na transformação de práticas sociais, econômicas e industriais.

Figura 16 - Nuvem de Palavra referente à categoria “*Technology*”



Fonte: Autora (2024).

A categoria *“Design & Fashion”* é marcada por termos como *“design”, “house”, “fashion”* e *“architect”*, que refletem uma conexão com questões relacionadas a projetos arquitetônicos de casas, edifícios e espaços, além de tendências no universo da moda. Esses termos destacam a relevância desse nicho de mercado, abrangendo não apenas a estética e funcionalidade de construções, mas também o impacto que estes trazem. A presença dessas palavras sugere a junção entre criatividade, inovação e funcionalidade, evidenciando por exemplo, como o *design* e a moda influenciam a percepção de valor e a experiência dos consumidores em diferentes contextos.

Figura 17 - Nuvem de Palavra referente à categoria “*Design & Fashion*”

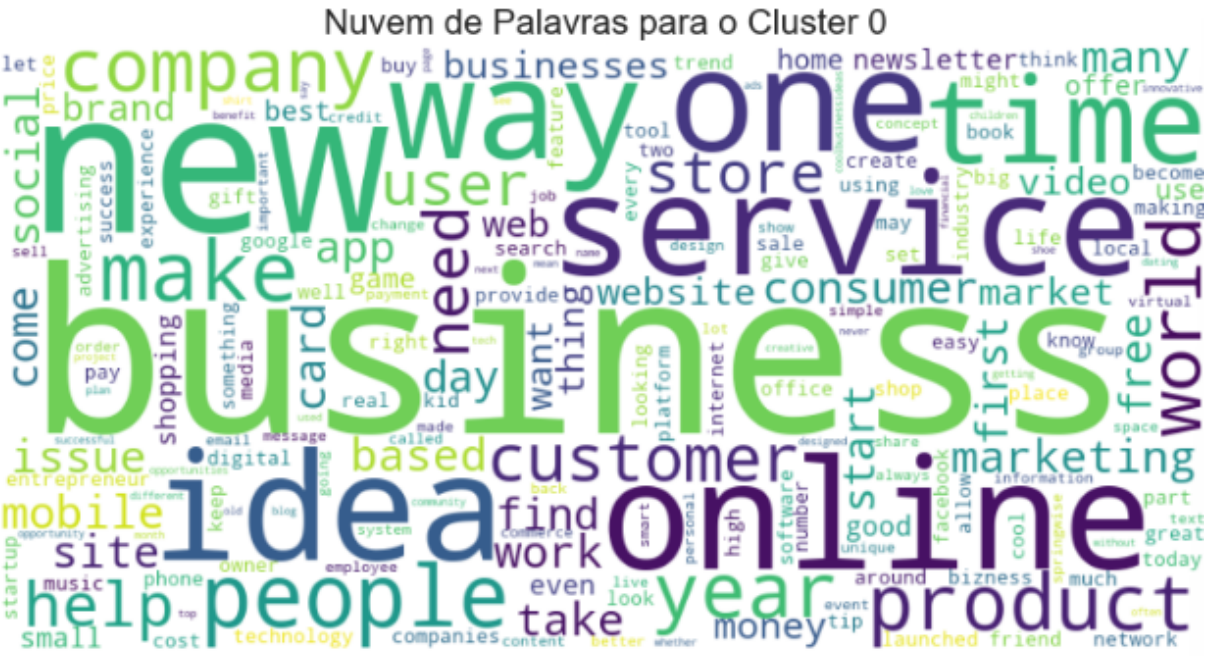
Fonte: Autora (2024).

4.2.2 Visualização de dados - *clusters*

Para uma análise mais detalhada, foram geradas nuvens de palavras e gráficos de barras para os cinco *clusters* definidos. A seguir, destacam-se os principais *insights*:

A Figura 18 apresenta o *cluster* 0, constituído de 1.899 ideias. Relacionado ao mundo corporativo, com termos como *business*, *service* e *company*. As categorias predominantes incluem ***"Services"*** e ***"Retail & Ecommerce"***.

Figura 18 - Nuvem de Palavra referente ao *Cluster* 0



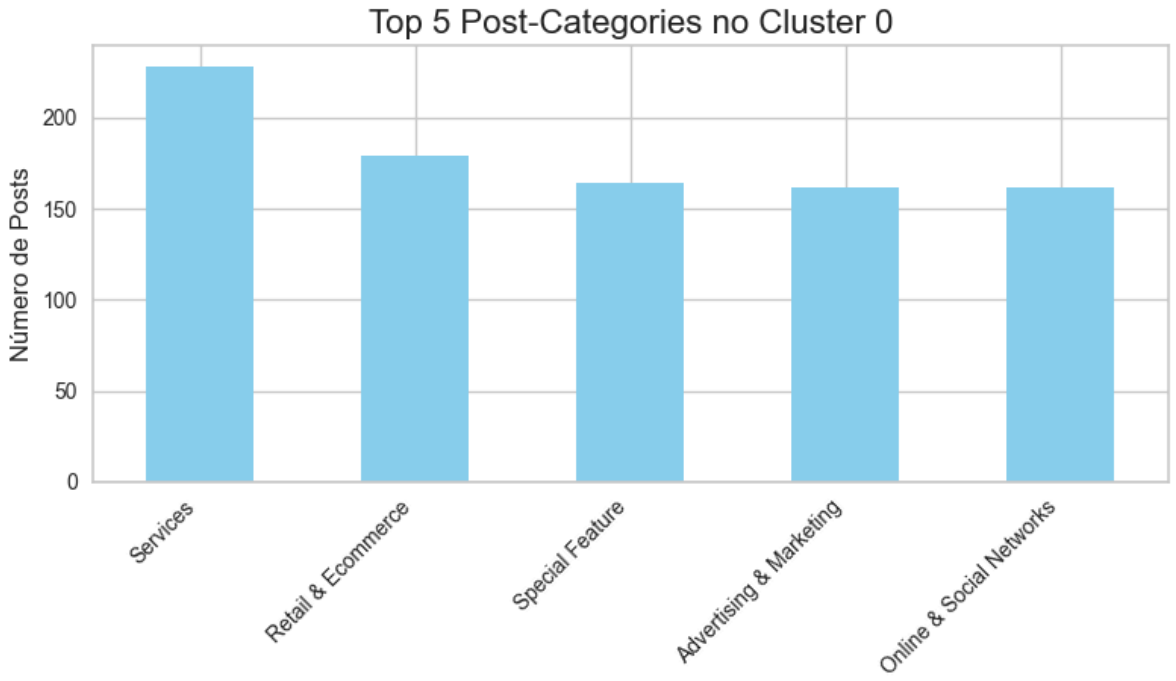
Fonte: Autora (2024).

Além disso, foram gerados gráficos de barras que explicitam as cinco categorias com maior número de ideias presentes em cada *cluster*. O objetivo desta visualização é extrair a correlação entre os termos-chave e as categorias.

A Figura 19 apresenta o *Cluster* 0 e, em consonância com a nuvem de palavras, destaca "*Services*" como a categoria líder, com mais de 200 ideias. Em seguida, estão as categorias *"Retail & Ecommerce"*, *"Special Feature"*, *"Advertising & Marketing"* e *"Online & Social Networks"*. Todas essas categorias permeiam o mundo corporativo ou empresarial de alguma maneira e são igualmente importantes para a constituição de um ecossistema de negócios dinâmico e diversificado.

Pode-se ainda inferir que essas categorias refletem a sinergia entre setores tradicionais e digitais, destacando como diferentes áreas colaboram entre si para criar um ambiente de negócios robusto e inovador. Essa análise permite compreender melhor as interações e a relevância de cada categoria no contexto geral dos *clusters*, fornecendo *insights* valiosos para futuras estratégias de negócios e inovação.

Figura 19 - Gráfico Top 5 Categorias *Cluster* 0



Fonte: Autora (2024).

O *Cluster* 1 destaca-se como o mais populoso entre os demais, com um total de 2.438 ideias. Ele inclui termos variados como "*food*", "*product*", "*water*", "*plastic*" e "*coffee*", conforme apresentado na Figura 20, indicando que as ideias presentes no *dataset* abrangem temas ligados a bens de consumo, sustentabilidade, necessidades cotidianas e outros aspectos correlatos.

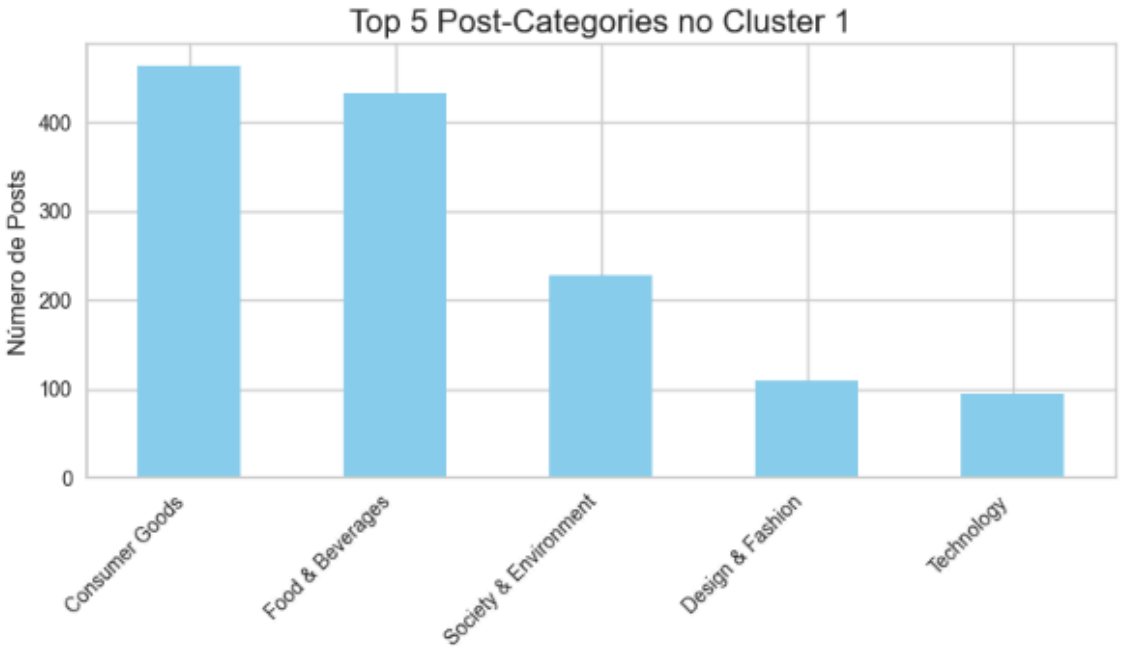
Corroborando a análise da nuvem de palavras, a Figura 21 evidencia as cinco principais categorias identificadas no *Cluster* 1: "*Consumer Goods*", "*Foods & Beverages*", "*Society & Environment*", "*Design & Fashion*" e "*Technology*", sendo que as duas primeiras contam com mais de 400 ideias cada. Isso pode ser um indicativo de que a diversidade de tópicos abordados permeia diferentes esferas, mas trata de contextos similares.

Por exemplo, ao considerar produtos de consumo, pode-se observar uma intersecção entre diferentes abordagens para o mesmo tema. Tal diversidade reflete a grande quantidade de ideias que se relacionam neste *cluster*, demonstrando como diferentes áreas colaboram para formar um ecossistema abrangente e interconectado.

Figura 20 - Nuvem de Palavra referente ao *Cluster* 1

Fonte: Autora (2024).

Figura 21 - Gráfico Top 5 Categorias *Cluster* 1



Fonte: Autora (2024).

O *Cluster* 2 apresenta um conjunto de 1.699 ideias. Na Figura 22, é apresentada a nuvem de palavras referente a este *cluster*, onde se encontram termos como *"design"*, *"home"*, *"space"* e *"project"*, entre outros, que remetem a conceitos de estética, organização e planejamento. Esses termos sugerem que as

ideias contidas no *Cluster* 2 se correlacionam com as áreas de design de interiores, arquitetura e decoração.

Figura 22 - Nuvem de Palavra referente ao *Cluster* 2

Fonte: Autora (2024).

A Figura 23 a seguir apresenta as cinco categorias com maior número de ideias dentro do *Cluster* 2. Observa-se que as categorias mais abrangentes são “*Design & Fashion*”, “*Homes & Living*”, “*Society & Environment*”, “*Travel & Transport*” e “*Consumer & Goods*”. Essas categorias destacam o escopo temático do *cluster*, que permeia desde o design e moda até questões relacionadas à sustentabilidade e responsabilidade social. Além disso, a presença de “*Travel & Transport*” e “*Consumer & Goods*” sugere conexões ligadas à mobilidade e ao turismo, sugerindo interseções entre as ideias abordadas.

Figura 23- Gráfico Top 5 Categorias *Cluster* 2

Fonte: Autora (2024).

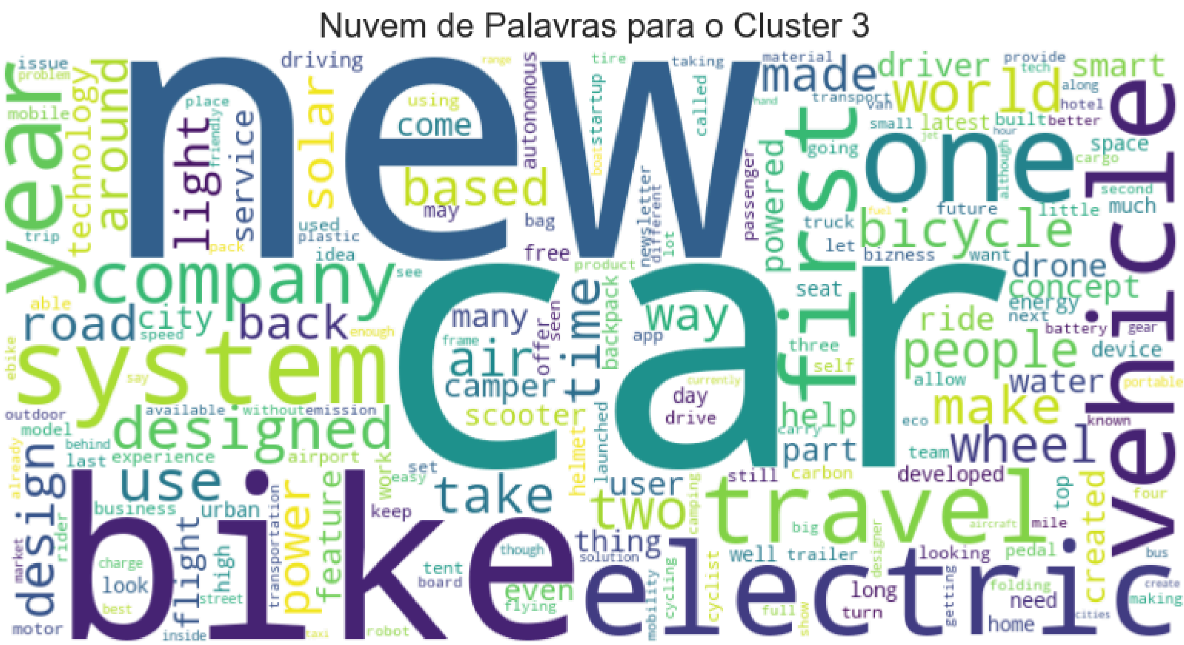
O *Cluster* 3, com 1.403 ideias, apresentado na Figura 24, contém termos como *"car", "bike"* e *"vehicle"*, indicando uma forte presença de ideias relacionadas à indústria automobilística. A nuvem de palavras sugere a presença de ideias ligadas ao campo da mobilidade e transporte, sejam motorizados ou não. Além disso, abrange palavras que remetem a soluções sustentáveis, veículos autônomos, infraestrutura, entre outros pontos relevantes.

Em consonância com esses termos, as cinco categorias que mais se destacaram neste *cluster* foram *"Travel & Transport"*, *"Consumer Goods*", *"Technology"* e *"Society & Environment"*, conforme ilustrado na Figura 25, sendo que *"Travel & Transport"* lidera com quase 600 ideias. Isso indica que este *cluster* apresenta uma interligação entre o recorte temático sugerido pela categoria e temas que incluem aspectos tecnológicos, sociais e ambientais.

A predominância das categorias *"Travel & Transport"* e *"Consumer Goods"* reforça a conexão com produtos e serviços relacionados ao transporte e à experiência de deslocamento. A presença da categoria *"Technology"* destaca a relevância de inovações, como veículos elétricos e autônomos, enquanto *"Society & Environment"* sublinha preocupações com sustentabilidade, impacto ambiental e

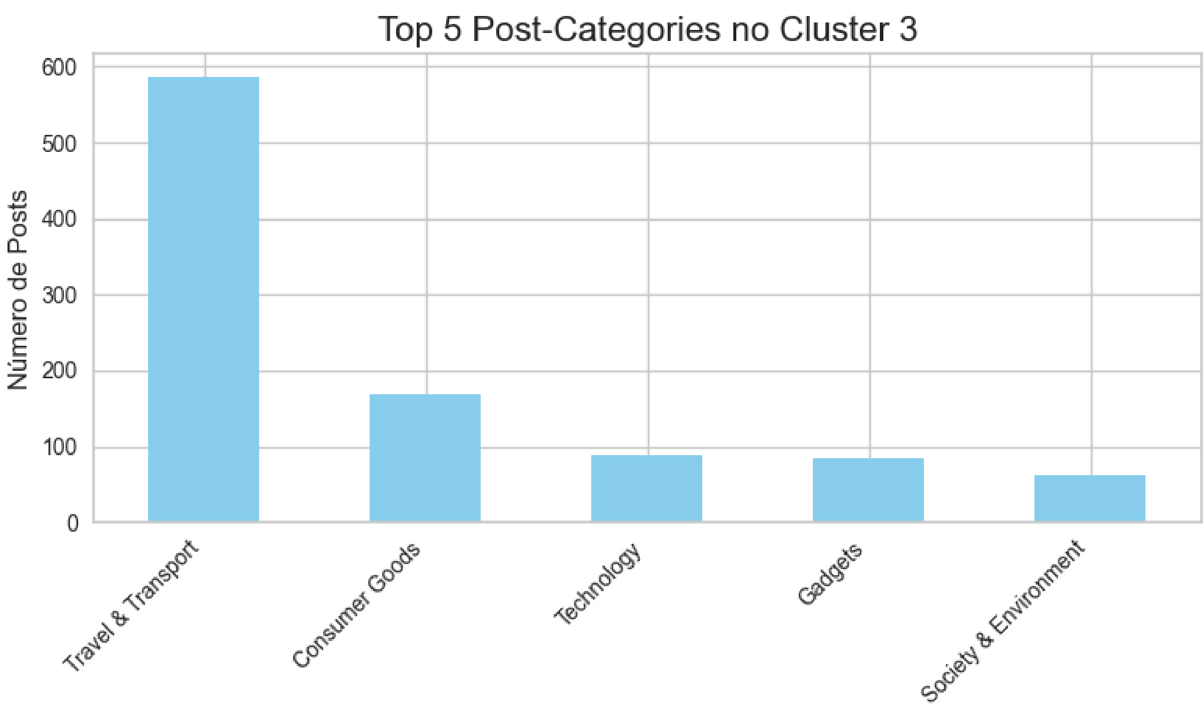
qualidade de vida. Dessa forma, o *cluster* reflete tendências que integram diversos fatores atuais e relevantes, evidenciando a sinergia entre diferentes áreas e a importância de abordagens multidisciplinares.

Figura 24 - Nuvem de Palavra referente ao *Cluster* 3



Fonte: Autora (2024).

Figura 25 - Gráfico Top 5 Categorias *Cluster* 3



Fonte: Autora (2024).

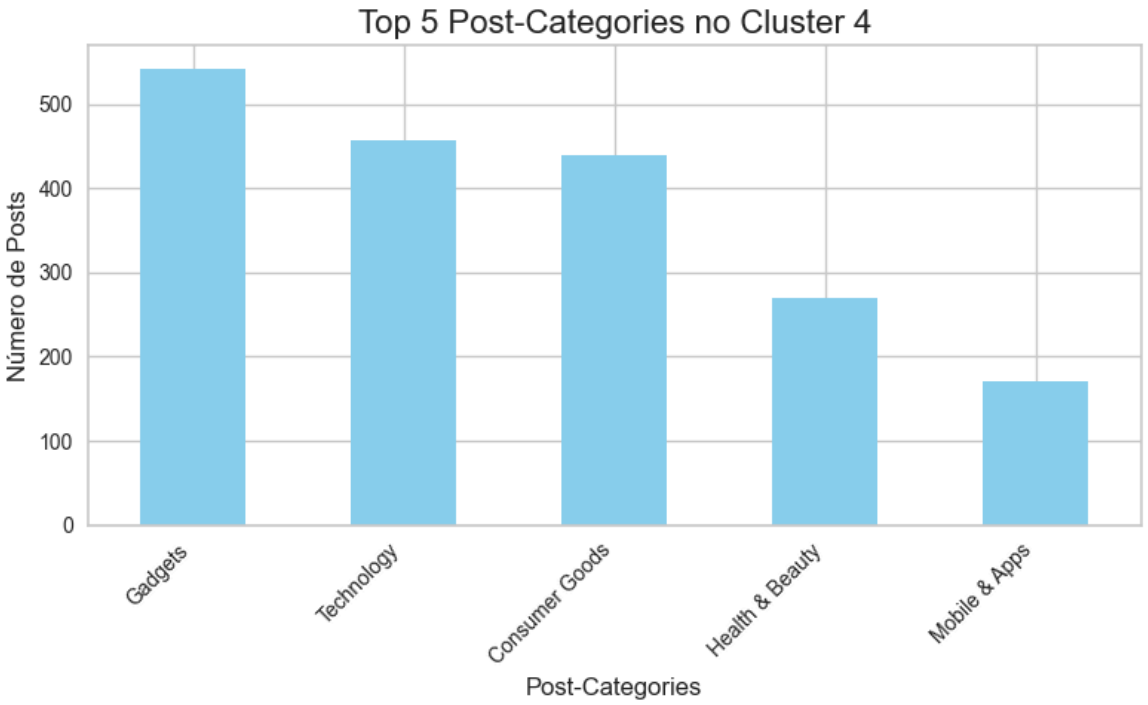
Por fim, o *Cluster* 4, ilustrado na Figura 26, conta com 1.266 ideias, sendo o menor entre os *clusters*. Ele inclui termos como *“smart”, “device”, “technology”, “phone”* e outros que indicam uma forte correlação com o universo da tecnologia e dos aparelhos eletrônicos.

Figura 26 - Nuvem de Palavra referente ao *Cluster* 4

Fonte: Autora (2024).

O grafo apresentado na Figura 27 revela as categorias predominantes neste *cluster*, destacando “*Gadgets*”, com mais de 500 ideias, seguida por “*Technology*”, com pouco mais de 450 ideias, e “*Consumer Goods*”, com mais de 400. As categorias *“Healthy & Beauty”* e *“Mobile & Apps”* aparecem em menor quantidade. Esses dados corroboram a análise anterior, que aponta que este *cluster* está centrado em torno de tópicos relacionados a produtos e dispositivos tecnológicos voltados para o consumidor final.

Figura 27 - Gráfico Top 5 Categorias *Cluster 4*



Fonte: Autora (2024).

Por fim, visando explorar maiores possibilidades de *insights*, optou-se por realizar um recorte temporal no dataset na época da pandemia de COVID-19, de fevereiro de 2020 até maio de 2021, e extrair uma visualização, com o objetivo de identificar tendências ou inclinações entre as ideias resgatadas.

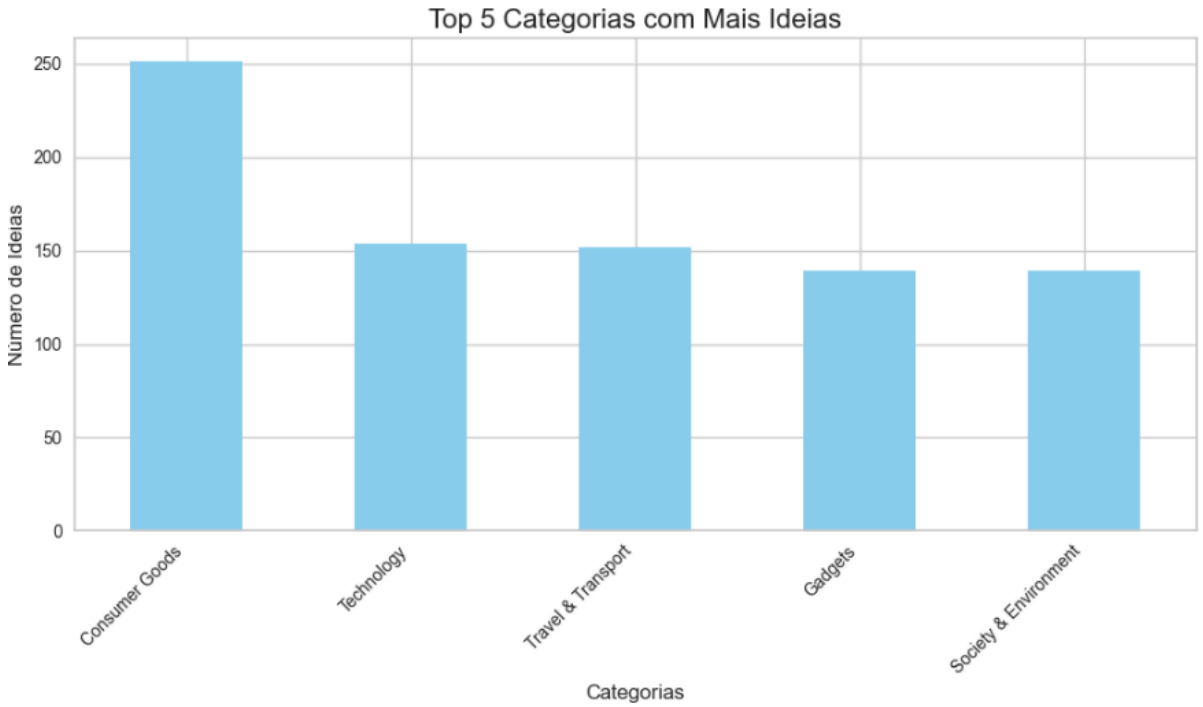
A Figura 28 apresenta termos como *“home”, “time”, “company”, “system”,* entre outras palavras que refletem as experiências e adaptações dos indivíduos durante o período de distanciamento social. Esses termos sugerem uma concentração temática em torno de aspectos como trabalho remoto, reorganização do tempo, e a necessidade de sistemas que apoiem atividades a partir de casa, destacando a influência dessas necessidades nas ideias resgatadas.

Figura 28 - Recorte temporal + nuvem de palavras

Fonte: Autora (2024).

Adicionalmente, correlacionando com a Figura 29, a análise revelou a predominância de tendências ligadas a soluções tecnológicas voltadas à conectividade, produtividade doméstica e ao bem-estar, indicando como a pandemia influenciou a priorização de temas como meio ambiente, estilo de vida, meios de transporte, etc. Essas inclinações ressaltam o papel de eventos globais na formação de tendências, demonstrando a relevância de métodos analíticos que capturem nuances temporais para a gestão de ideias.

Figura 29 - Recorte temporal + Top 5 categorias



Fonte: Autora (2024).

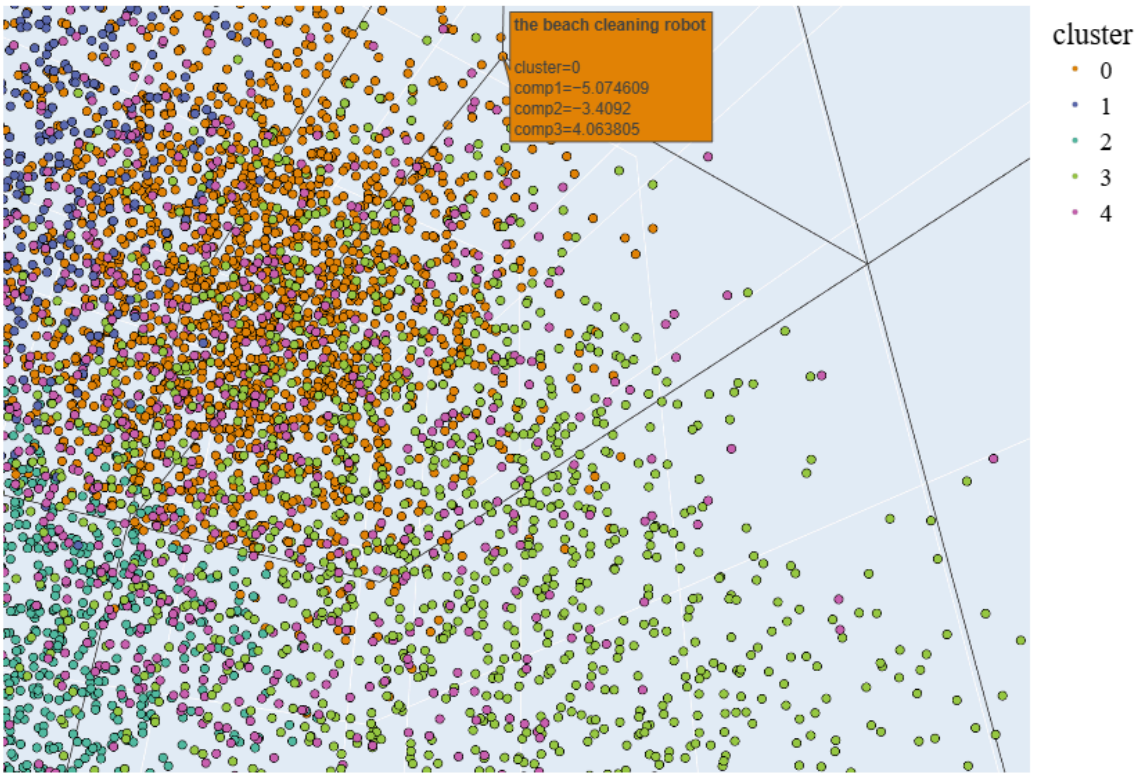
4.2.3 Busca semântica

Para extrair insights mais profundos do *dataset*, foram realizadas buscas semânticas explorando diferentes aspectos dos dados, como padrões de similaridade entre ideias, relações contextuais entre categorias e a evolução temporal de certos conceitos.

A primeira análise baseou-se na combinação da visualização do PCA com a busca semântica, com o objetivo de entender melhor a composição dos dados e identificar *insights* relevantes.

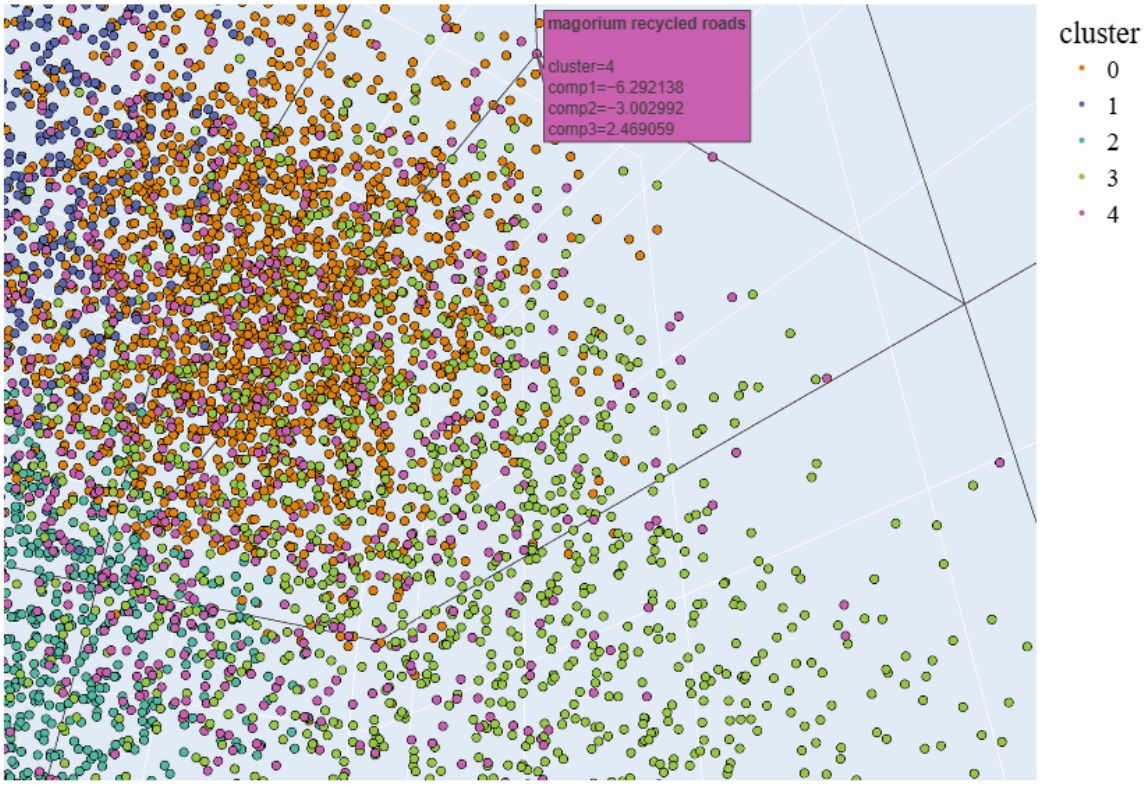
As Figuras 30, 31 e 32 apresentam ideias pertencentes a três *clusters* distintos, mas localizadas próximas ao "limite" entre os respectivos agrupamentos. Embora sejam classificadas em *clusters* diferentes, observa-se que essas ideias possuem temas inter-relacionados, evidenciando uma proximidade conceitual que transcende os limites definidos pela segmentação.

Figura 30 - Visualização PCA + Busca semântica *Cluster* 0



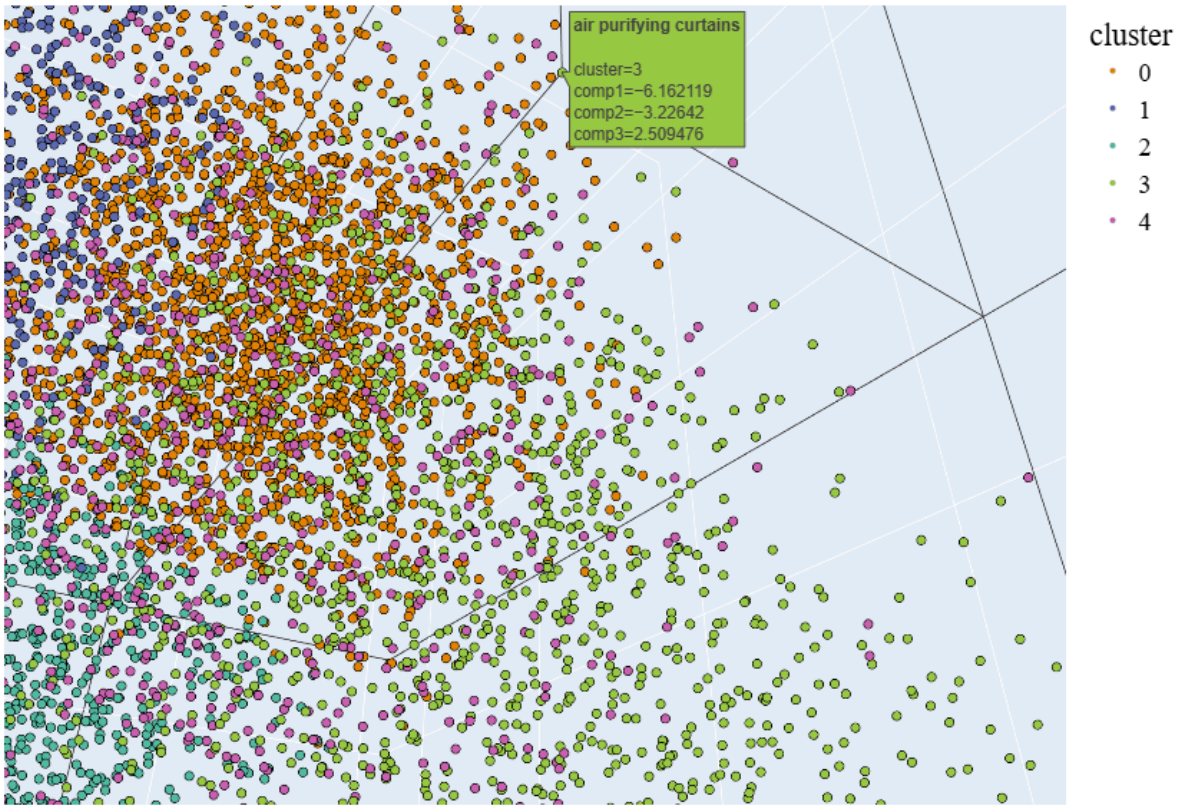
Fonte: Autora (2024).

Figura 31 - Visualização PCA + Busca semântica *Cluster* 4



Fonte: Autora (2024).

Figura 32 - Visualização PCA + Busca semântica *Cluster* 3



Fonte: Autora (2024).

Exemplos como os títulos *“The Beach Cleaning Robot” (Cluster 0), “Megorium Recycled Roads” (Cluster 4)* e *“Air Purifying Curtains” (Cluster 3)* ilustram essa correlação. Essas ideias abordam questões de sustentabilidade e inovação tecnológica em diferentes contextos, como a limpeza ambiental, a reutilização de materiais em infraestrutura e a melhoria da qualidade do ar.

Essa análise demonstra que, apesar da divisão em *clusters*, há uma ligação entre os temas, indicando tendências compartilhadas entre os grupos. Isso ressalta a importância de uma análise global dos dados para a descoberta de *insights* mais profundos e reveladores.

A segunda análise consiste em identificar as cinco ideias mais similares de cada categoria. Esse processo permite observar a evolução das ideias, bem como identificar padrões e conexões temporais, promovendo uma visão mais clara das tendências e transformações do pensamento dentro de cada área.

O Quadro 3, a seguir, apresenta a relação dos dados retornados da categoria *"Business Trends"*.

Quadro 3 - Top 5 ideias mais similares Categoria "*Business Trends*"

Média de Similaridade	Título	Resumo	Data de publicação
0.513143391007244	5 steadily rising 2017 small business trends	Article contributed by anna johansson - any entrepreneur or business minded individual knows there are two things that are certain in business: taxes and change.Though there are timeless keys, like friendly customer service and the use of the web for business interactions, many trends will come and go with each passing year, some trends from last year have already gone out [...]	2017-10-13
0.5125283675618691	The most important elements of a business.	Establishing solid foundations there are many diverse aspects of business ownership you'll need to take into account for finest success. You can't just start buying and selling, trading, providing products, services or whatever else constitutes your business as simply as deciding to, well you can, but your success will take some severe perseverance if you approach it without preparation a business [...]	2018-02-14
0.5043996693474231	4 ways to improve the scalability of your business.	Every company needs to have a plan for its own growth, indeed a business owner who isn't looking to expand isn't likely to experience much sustained success. To put it even more bluntly, if you're not growing as a business, you're dying. With that in mind, today we're going to look at four ways that startups can improve their scalability [...]	2018-06-13
0.5005369690997534	4 industries that are projected to grow exponentially in 2020.	The new decade is nearly knocking on our door, what does business have in store for startup owners and aspiring entrepreneurs? We've watched markets rise and fall over the past 20 years and investors who got burned by the volatility know how important it is to financially forecast the years to come. Here's a look at four industries that [...]	2019-10-15
0.5053404357173655	Expert advice for young business owners.	Are you a young business owner looking to take your company to the next level? Depending on your industry and sector, it can seem daunting, but don't worry whether you're looking to manage a small business out of your home or want to start a major company similar in size to black tie moving columbus, we're here to help [...]	2023-04-14

Fonte: Autora (2024).

Baseado nos 5 retornos, pode-se perceber que há um ponto focal de estratégia de crescimento e adaptação das empresas nestas ideias. Olhando no viés evolutivo das publicações, pode-se observar uma transição entre o foco na fundamentação e previsões para uma abordagem mais prática, voltada à aplicação de estratégias de longo prazo e visão de crescimento sustentável. Ainda, as médias de similaridade apontam a coesão entre essas ideias.

Já o Quadro 4 abaixo apresenta a relação dos dados retornados da categoria "*Health & Beauty*".

Quadro 4 - Top 5 ideias mais similares Categoria "*Health & Beauty*"

Média de Similaridade	Título	Resumo	Data de publicação
0.38620067480425374	Can this wearable keep you less stressed…	You'd be forgiven for assuming that wearable tech is all about fitness trackers to help shed the pounds and keep you on track to reach your fitness goals. We've seen a steady rise in tech aimed at keeping our minds in shape just as much as our bodies. Stress tracking is one of the most wanted features from wearable tech, while major players like Apple [...]	2018-03-18
0.375641869763421	stretchy wearable tech	Wearable devices can be stuck onto the stick like plasters. Using a new method pioneered by researchers at Carnegie Mellon university, a combined effort from engineers at the school's soft machines lab and morphing matter lab. Electrodermis is a new way to apply electronics to the skin, whether it be for medical, fitness or lifestyle purposes. Where current wearable technology [...]	2019-07-09
0.37159563749747987	shirt tracks vitals	Although we've recently seen a number of electronic skin sensor patches that monitor the wearer's vital signs, the things do have their drawbacks. Scientists at MIT have therefore developed what may be a better alternative, in the form of a vitals monitoring shirt. Electronic skin patches typically take the form of a thin sheet of silicone with electronic components embedded inside [...]	2020-05-06
0.37086076849261246	the future of medical wearables	wearable health monitors are everywhere from fitbits for the health conscious to continuous glucose monitors for diabetics but most are limited in what they can tell us and there are issues around accuracy calibration and reliability researchers in sweden are working to change that the technology developed by scientists at kth royal institute of technology in sweden employs multi purpose electrochemical sensors	2019-06-01

0.369522859835243	get to know your skin conditions today in an instant	have you ever looked at your dull lifeless skin and wondered how it got that way and how you could bring it back to its former glory cracking the code can be tricky eating well getting plenty of water and using the right skin care products all play a part but sometimes despite our best efforts our skin just	2018-03-23

Fonte: Autora (2024).

Ao analisar as ideias acima, percebe-se que estas permeiam a exploração de tecnologias vestíveis e dispositivos inovadores destinados à melhoria do bem-estar físico e mental dos indivíduos. Percebe-se diversidade nas abordagens, que vão desde o monitoramento dos sinais vitais e estado de saúde até ferramentas para gerenciamento do estresse e cuidados com a pele. O recorte temático apresentado por estas ideias apontam uma uma integração entre tecnologia e academia voltada à saúde, buscando promover aplicações práticas para suprir demandas específicas bem como a aderência das pessoas a novas formas de cuidados pessoais.

Por fim, o Quadro 5 traz a relação dos dados retornados da categoria "*Finance & Startups*".

Quadro 5 - Top 5 ideias mais similares Categoria "*Finance & Startups*"

Média de Similaridade	Título	Resumo	Data de publicação
0.5439794743061066	Tackling the challenges of raising capital for your start up.	A brilliant idea can be the start of a successful business, but it's not the only thing you need. One of the main reasons that many startups fail is because they are under capitalized, despite the popular stories of entrepreneurs launching a multi million dollar concern in their bedroom with a nickel and a broken typewriter, it is almost impossible to [...]	2016-01-09
0.5506977158784866	Three tips for the struggling entrepreneur…	So you're running your own business, you've left cubicle life behind for good to be your own boss and it's been an exciting ride. But now you're in trouble, your seemingly loyal customer base is drying up, and your operating expenses are only growing. You had a leg up on your competition, though now it seems the tables have turned [...]	2016-05-31
0.5777632713317871	How to get hold of the cash you already have for new business, ideas and startups…	As a new business owner, you might not have every business practice totally mastered, for example, your payroll process might need some retooling, and your pos system might benefit from a few upgrades.	2016-12-13

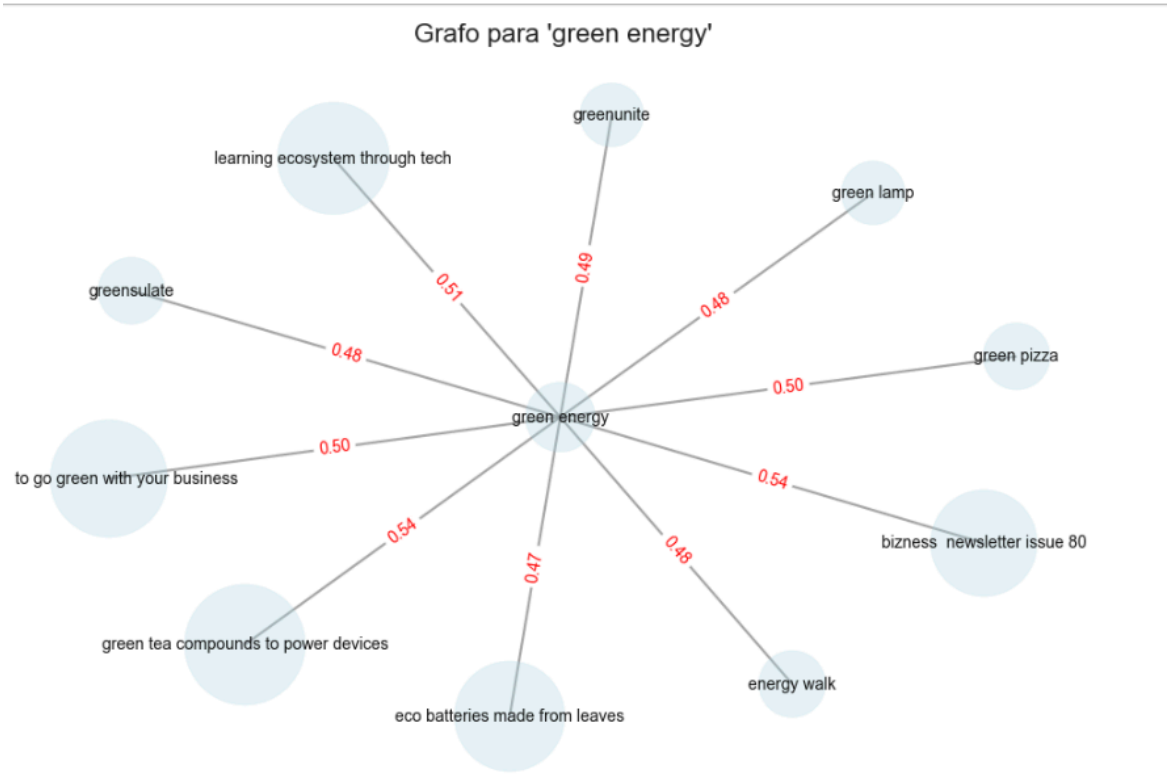
		The most finicky of all your business practices, and the one that new entrepreneurs most often get wrong, is your cash flow, namely, balancing the [...]	
0.5699953144788742	How to finance a new business.	When you're trying to start a new business, one of the hardest parts of doing so is getting the funding together to support it, after all, where do you start when your own personal bank account and savings run dry? There are plenty of ways to do it, from bank loans to crowdfunding, to credit cards, funding your business [...]	2019-06-05
0.5435151261091232	Getting ready to launch a new idea innovation, or start up a new business	Launching or starting a new business is a very exciting time, there is so much happening and going on, even behind the scenes, that it can be hard to contain your excitement and anticipation. When you are looking at launching a new, cool or innovative idea, there are areas that you will want to focus on to ensure that you [...]	2021-07-31

Fonte: Autora (2024).

Ademais, foi realizada uma busca semântica de termos-chave escolhidos aleatoriamente, visando compreender a correlação entre os resultados obtidos. A Figura 33 apresenta o grafo gerado a partir dos outputs relacionados ao termo *"Green energy"*. Títulos como *"Green tea compounds to power devices"*, *"Eco batteries made from leaves"* e *"Green lamp"* indicam uma tendência à exploração de fontes sustentáveis e alternativas para a produção de energia. Esses títulos sugerem uma sinergia entre sustentabilidade e tecnologia no desenvolvimento de soluções ecológicas e eficientes.

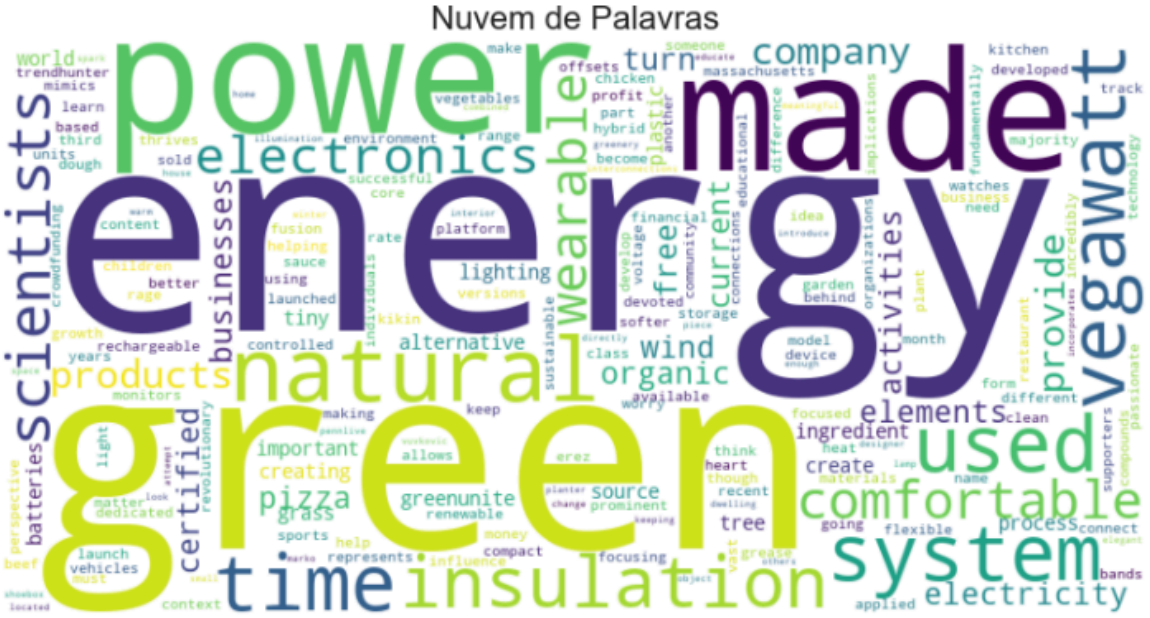
Para corroborar a análise do grafo, foi gerada uma nuvem de palavras a partir do retorno obtido com a busca semântica. A Figura 34 apresenta termos como *"vegawatt", "natural", "scientists",* além de *"green" e "energy"*, que reforçam a associação entre inovação científica e práticas sustentáveis. Esses termos sugerem um foco em soluções baseadas em recursos naturais, integrando ciência, tecnologia e sociedade para promover o desenvolvimento de um futuro mais sustentável.

Figura 33 - Grafo “*Green energy*”



Fonte: Autora (2024).

Figura 34 - Nuvem de Palavras “*Green energy*”

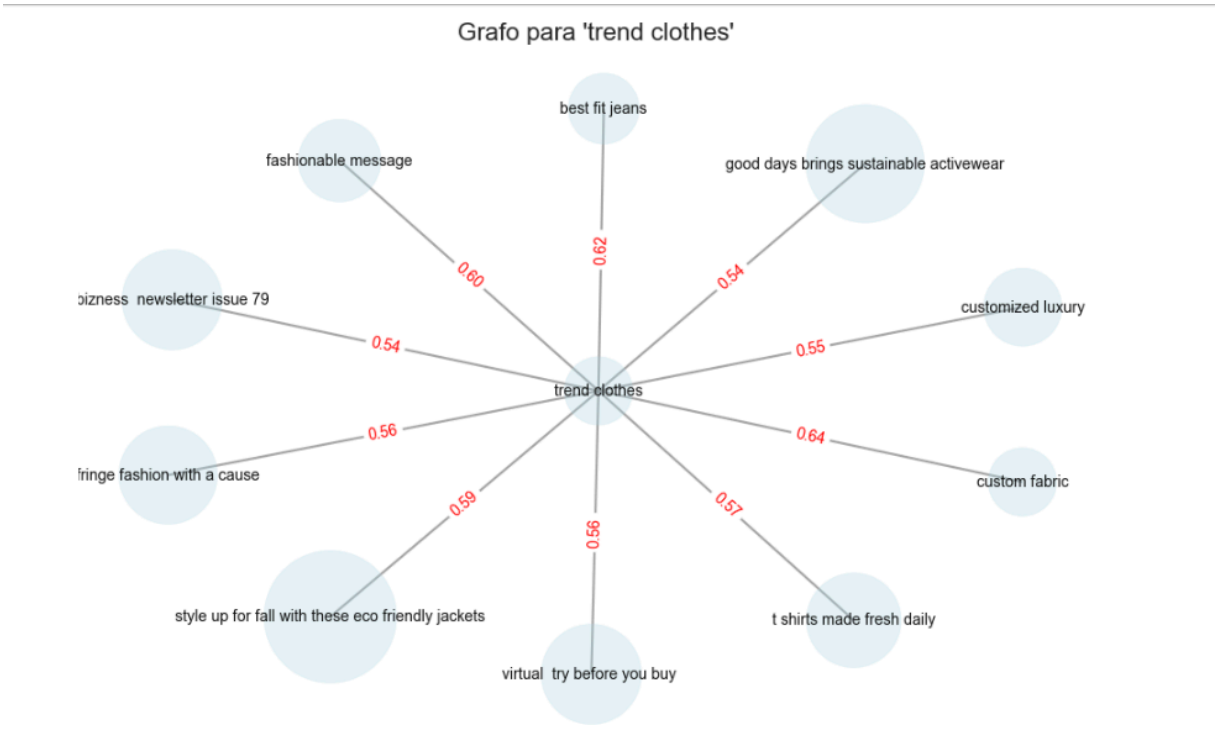


Fonte: Autora (2024).

A Figura 35 apresenta o grafo gerado a partir das saídas relacionadas ao termo “*Trend Clothes*”. Títulos como “*Good days brings sustainable activewear*”, “*Fashionable message*” e “*Style up for fall with these eco friendly jackets*” apontam uma inclinação ao uso de roupas de modo sustentável e consciente. Ainda, títulos como “*best fit jeans*” e “*Virtual try before you buy*” apontam que há um interesse por parte dos consumidores na compra assertiva dos produtos de moda.

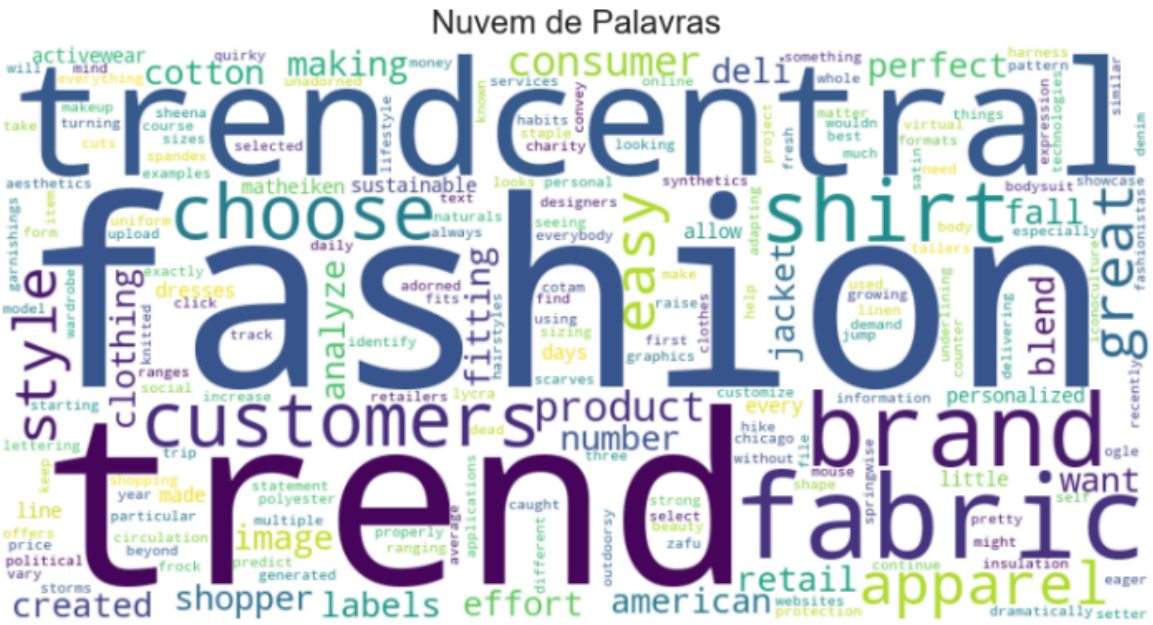
Em colaboração com a análise do grafo, gerou-se uma nuvem de palavras do retorno obtido com a busca semântica. A Figura 36 apresenta termos como “*customer*”, “*fashion*”, “*trend*”, entre outros, que reforçam os pontos focais das ideias retornadas.

Figura 35 - Grafo “*Trend clothes*”



Fonte: Autora (2024).

Figura 36 - Nuvem de Palavras “*Trend clothes*”

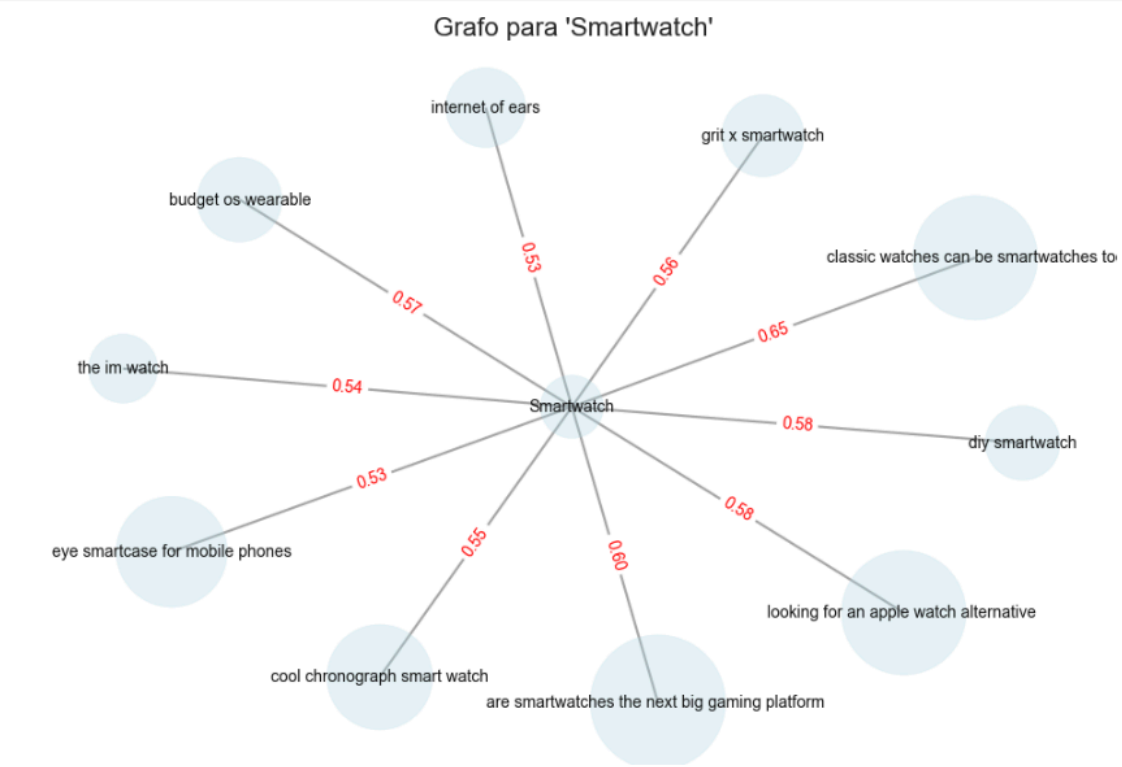


Fonte: Autora (2024).

Por fim, a Figura 37 apresenta o grafo gerado a partir do termo “*Smartwatch*”. Títulos como “*Gift x Smartwatch*”, “*Looking for an Apple watch alternative*” e “*Are smartwatches the next big gaming platform*” apontam um desenvolvimento contínuo deste nicho de mercado.

Interseccionando com a nuvem de palavras referente à mesma busca semântica conforme Figura 38, temos a presença de termos como “*Smartwatch*”, “*smartphone*”, “*wearable*”, “*device*”, “*bluetooth”*, entre outros que nos remetem à aparelhos tecnológico inteligentes e que permitem conectividade, portabilidade e até mesmo integração com outros dispositivos. Essa interseção não reflete apenas as tendências de consumo, mas também os avanços tecnológicos em diferentes áreas, que impulsionam o interesse por funcionalidades mais amplas e diversificadas.

Figura 37 - Grafo “*Smartwatch*”



Fonte: Autora (2024).

Figura 38 - Nuvem de Palavras “*Smartwatch*”

Fonte: Autora (2024).

4.2.4 Insights obtidos a partir das visualizações

A análise detalhada das ideias presentes no *dataset* revelou conexões temáticas significativas entre as diferentes categorias. Por meio de visualizações como nuvens de palavras, grafos e gráficos de barras, foi possível identificar a presença de temas amplamente alinhados aos debates contemporâneos, como sustentabilidade, inovação e tecnologia. Um ponto focal observado foi a combinação de ideias de natureza diversa com uma ou mais dessas três áreas, evidenciando a interseção entre domínios tradicionais e avanços modernos.

Exemplos interessantes incluem a integração da sustentabilidade com a moda, o uso da tecnologia para desenvolver soluções inovadoras na área da saúde, e a adoção de diferentes métodos de produção de energia “verde”, que unem, de forma eficaz, tecnologia e sustentabilidade. Essas combinações não apenas ilustram a diversidade temática do *dataset*, mas também refletem a crescente convergência de disciplinas na busca por soluções abrangentes para os desafios globais contemporâneos.

Neste sentido, as nuvens de palavras geradas para cada *cluster* apontaram focos temáticos distintos; no entanto, observou-se a recorrência de termos que remetem a valores compartilhados, como responsabilidade ambiental, eficiência tecnológica e impacto social positivo. Essa constatação reforça a ideia de que, apesar das especificidades de cada grupo, há uma interconexão entre os temas que sugere uma direção comum na busca por inovação sustentável e soluções de longo alcance.

Essa recorrência está alinhada com as tendências globais emergentes, onde a integração entre sustentabilidade, inovação e tecnologia tem se tornado cada vez mais central para a promoção de avanços econômicos, sociais e ambientais. Essa dinâmica ressalta a importância de analisar e fomentar iniciativas interdisciplinares que aproveitem essas sinergias, com o objetivo de gerar impactos positivos e ampliar o alcance de transformações significativas. Em última instância, essas interconexões oferecem uma base sólida para o desenvolvimento de soluções integradas e sustentáveis, que são capazes de enfrentar os desafios complexos do século XXI.

4.2.5 Implicações para auxílio da gestão de ideias nas Organizações

As análises realizadas neste estudo revelam oportunidades estratégicas significativas para as organizações no aproveitamento de fontes de dados e ideias disponíveis online. A capacidade de extrair *insights* valiosos dessas informações pode proporcionar vantagens competitivas, permitindo que empresas inovem de forma mais assertiva e eficaz, respondendo de maneira ágil às crescentes demandas do mercado.

Uma das principais descobertas refere-se à sinergia interdisciplinar evidenciada nas visualizações analisadas. Combinações como sustentabilidade e tecnologia demonstram grande potencial para gerar inovações tanto em produtos quanto em processos, resultando em soluções mais alinhadas às prioridades globais, como eficiência ambiental, funcionalidade e impacto social positivo. Ao explorar essas interseções, as organizações podem desenvolver produtos diferenciados e altamente aderentes às expectativas de consumidores e stakeholders, promovendo um futuro mais sustentável e tecnologicamente avançado.

Além disso, as análises temporais oferecem uma compreensão mais aprofundada sobre a evolução de temas e tendências, proporcionando uma perspectiva estratégica essencial para as empresas. Isso permite que as organizações não apenas antecipem movimentos de mercado e identifiquem oportunidades emergentes, mas também adaptem suas estratégias para explorar novos nichos e conquistar segmentos de mercado ainda inexplorados. Essa visão dinâmica e proativa é crucial para a continuidade do sucesso e para a renovação da relevância organizacional.

Por fim, a compreensão clara das preferências e necessidades dos consumidores dentro do nicho de mercado visado é um fator determinante para o sucesso das iniciativas. Essa proximidade possibilita a criação de soluções que atendam diretamente às “dores” dos clientes, alinhando as ofertas com as expectativas do público-alvo. Consequentemente, as organizações podem não apenas aumentar o engajamento e a lealdade dos consumidores, mas também fortalecer sua posição competitiva no mercado, promovendo iniciativas mais significativas, impactantes e sustentáveis.

4.2.6 Limitações do estudo e trabalhos futuros

Apesar de permitir a extração de *insights* a partir da análise de ideias, o *dataset* utilizado neste estudo apresenta algumas limitações significativas. Primeiramente, ele não abrange a totalidade do ecossistema global de áreas temáticas, limitando a visão sobre as diversas possibilidades existentes em termos de inovação. Além disso, pode conter vieses inerentes à plataforma de onde os dados foram extraídos, principalmente no que diz respeito à seleção das ideias publicadas, que podem não refletir toda a diversidade de perspectivas e abordagens possíveis. Outra limitação importante foi a análise restrita às descrições das ideias ('entry-content'), o que desconsidera outros elementos textuais disponíveis, como comentários, tags ou informações contextuais adicionais, que poderiam enriquecer a interpretação e proporcionar um entendimento mais amplo dos temas tratados.

Para trabalhos futuros, recomenda-se a implementação de um banco de dados estruturado, que garantisse a persistência dos dados e facilitasse o acesso e a reutilização. A criação de um repositório mais robusto também possibilitaria a realização de análises longitudinais, permitindo a observação das tendências ao longo do tempo. Além disso, explorar diferentes modelos de *embeddings* e técnicas avançadas de clusterização poderia aprimorar a qualidade do processamento e da análise, proporcionando uma segmentação mais precisa das ideias. A inclusão de novas abordagens, como técnicas de visualização de dados mais avançadas e a aplicação de busca semântica refinada, contribuiria significativamente para uma compreensão mais profunda e abrangente do conjunto de ideias, permitindo a identificação de padrões complexos e relações sutis entre as categorias.

Por fim, embora este estudo tenha apresentado limitações evidentes, ele demonstrou a relevância de análises estruturadas para identificar tendências temáticas e interconexões entre áreas de conhecimento, essencial para a compreensão das dinâmicas de inovação. Ao mesmo tempo, o estudo abriu caminho para futuras investigações, enfatizando a importância de métodos robustos e abrangentes na análise de grandes volumes de dados. Assim, os resultados alcançados não apenas oferecem uma base para futuros avanços técnicos, mas também reforçam o papel estratégico da análise de dados na promoção de inovações alinhadas às demandas globais, permitindo que as organizações se adaptem de maneira mais eficaz às transformações do mercado e da sociedade.

5 CONSIDERAÇÕES FINAIS

Este trabalho investigou o potencial da análise de ideias para a gestão estratégica e a inovação organizacional, utilizando um conjunto de técnicas avançadas de análise de dados. A partir da coleta e análise de sugestões de diversas categorias, extraídas de uma plataforma de ideias, o estudo foi capaz de identificar padrões e tendências emergentes que são fundamentais para entender como diferentes temas, muitas vezes classificados de forma distinta, se conectam em um contexto mais amplo. O uso de ferramentas como BERT, K-means, PCA e LLM, combinadas de forma integrada, permitiu a vetorização, clusterização e redução de dimensionalidade dos dados, resultando em uma análise aprofundada das ideias.

Uma das principais conclusões deste estudo é a identificação de tendências globais e interdisciplinares, que se concentram em questões de relevância mundial, como sustentabilidade, inovação tecnológica e eficiência social. Esses temas, quando analisados de diferentes perspectivas, destacam a importância da interconexão entre áreas de conhecimento, o que proporciona uma base sólida para o desenvolvimento de soluções inovadoras e a geração de impacto positivo.

A pesquisa também demonstrou que as ideias, embora agrupadas em categorias específicas, podem abordar temas semelhantes sob ângulos distintos, o que amplia a compreensão sobre as inter-relações existentes entre diferentes áreas de interesse. Esse entendimento contribui diretamente para a gestão de ideias, permitindo que gestores e empreendedores identifiquem pontos focais que podem ser explorados para aprimorar produtos e processos. A implementação do método de vetorização e clusterização, além de garantir a precisão na extração de dados, demonstrou ser eficaz no processamento de grandes volumes de informações e na identificação de insights relevantes.

Além disso, a análise automatizada do conteúdo não só possibilitou a extração de *insights* significativos, mas também levantou reflexões sobre a aplicabilidade dessas técnicas na gestão de ideias, destacando seu potencial em apoiar decisões estratégicas e no planejamento organizacional. A pesquisa também foi bem-sucedida ao atingir seus objetivos específicos, demonstrando como as diferentes abordagens analíticas, como o uso de LLM e agrupamento de dados,

podem ser aplicadas de maneira eficiente para a gestão de grandes volumes de ideias.

Por fim, este estudo apresenta implicações práticas para empreendedores e gestores, fornecendo uma metodologia robusta que pode ser utilizada para analisar tendências e identificar oportunidades no mercado. O método desenvolvido neste trabalho não apenas contribui para a gestão de ideias, mas também oferece uma base sólida para futuras investigações, que poderão replicar ou adaptar as técnicas propostas em diferentes contextos de pesquisa. Dessa forma, a análise e visualização de ideias se consolida como uma ferramenta estratégica, capaz de gerar *insights* decisivos para a inovação e o sucesso organizacional.

REFERÊNCIAS

ABDURRAHMAN, L; MULYANA, T. Parallel Construction of Information Technology Value Model: Design-Science Research Methodology. **In: 8th International Conference on Information and Communication Technology (ICoICT)**, Yogyakarta, Indonesia, 2020.

ADIWARDANA, Daniel *et al*. Towards a Human-like Open-Domain Chatbot. **Arxiv**, [S.L.], v. [], n. [], p. 1-38, 27 jan. 2020. ArXiv. http://dx.doi.org/10.48550/ARXIV.2001.09977.

AHMED, A.; SHAHBA, R. Abou; GHAYAD, Ibrahime; ATTIA, E.; HUSSEIN, W.. ELECTROCHEMICAL BEHAVIOR OF STEEL ALLOYS AS AFFECTED BY PHOSPHORIC ACID. **Al-Azhar Bulletin Of Science**, [S.L.], v. 19, n. 1-, p. 153-167, 1 jun. 2008. Al-Azhar University. http://dx.doi.org/10.21608/absb.2008.9000. Disponível em: https://absb.researchcommons.org/journal/vol19/iss1/16/. Acesso em: 17 abr. 2024.

ARCAS, Blaise Agüera y. Do Large Language Models Understand Us? **Daedalus**, [S.L.], v. 151, n. 2, p. 183-197, 2022. MIT Press. http://dx.doi.org/10.1162/daed_a_01909. Disponível em: https://doi.org/10.1162/daed_a_01909 . Acesso em: 01 maio 2024.

AUFFARTH, Ben. **Generative AI with LangChain**: build large language model (llm) apps with python, chatgpt, and other llms. Birmingham: Packt Publishing Ltd, 2023. 336 p.

BAILEY, Brian P.; HORVITZ, Eric. What's your idea? **Proceedings Of The Sigchi Conference On Human Factors In Computing Systems**, [S.L.], v. 3, n. [], p. 2065-2074, 10 abr. 2010. ACM. http://dx.doi.org/10.1145/1753326.1753641. Disponível em: https://dl.acm.org/doi/10.1145/1753326.1753641 . Acesso em: 01 maio 2024.

BAKKER, Han; BOERSMA, Kees; OREEL, Sytse. Creativity (Ideas) Management in Industrial R&D Organizations: a crea ‣political process model and an empirical

illustration of corus rd&t. **Creativity And Innovation Management**, [S.L.], v. 15, n. 3, p. 296-309, set. 2006. Wiley. http://dx.doi.org/10.1111/j.1467-8691.2006.00397.x.

BAKKER, Hendrik Jan. Idea management: unraveling creative processes in three professional organizations. 2010.

BARBIERI, J. C.; ÁLVARES, A. C. T.; CAJAZEIRA, J. E. R. Gestão de Ideias para inovação contínua. Porto Alegre: **Bookman**, 2009.

BAREGHEH, A.; ROWLEY, J.; SAMBROOK, S. Towards a multidisciplinary definition of innovation. **Management Decision**, United Kingdom, v. 47, n .8, p.1323-1339, 2009.

BAYAZIT, Nigan. Investigating Design: A Review of Forty Years of Design Research. **Design Issues.** [S.l], p. 16-29. 1 jan. 2004. Disponível em: https://www.jstor.org/stable/1511952 . Acesso em: 12 abr. 2024.

BERGER, Jonah; PACKARD, Grant. Using natural language processing to understand people and culture. **American Psychologist**, v. 77, n. 4, p. 525, 2022.

BOEDDRICH, Heinz-Juergen. Ideas in the Workplace: a new approach towards organizing the fuzzy front end of the innovation process. **Creativity And Innovation Management**, [S.L.], v. 13, n. 4, p. 274-285, 8 nov. 2004. Wiley. http://dx.doi.org/10.1111/j.0963-1690.2004.00316.x. Disponível em: https://doi.org/10.1111/j.0963-1690.2004.00316.x . Acesso em: 15 abr. 2024.

BREM, A.; VOIGT, K. I. **Innovation management in emerging technology ventures** – the concept of an integrated idea management. International Journal of Technology, Policy and Management, v. 7, n. 3, p. 304-321, 2007.

BREM, Alexander; GIONES, Ferran; WERLE, Marcel. The AI digital revolution in innovation: A conceptual framework of artificial intelligence technologies for the management of innovation. **IEEE Transactions on Engineering Management**, v. 70, n. 2, p. 770-776, 2021.

Brownlee, J.: Deep learning & artificial neural networks in deep learning (2019)

CARBONE, F.; CONTRERAS, J.; HERNÁNDEZ, J. Z.; GOMEZ-PEREZ, J. M. Open Innovation in an Enterprise 3.0 framework: Three case studies. **Expert Systems with Applications**. 2012.

CASSIANO, K. M.; SOUZA, R. C. **Análise de séries temporais usando Análise Espectral Singular (SSA) e clusterização de suas componentes baseada em densidade. [recurso eletrônico]**. *[S. l.: s. n.]*. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&db=cat06910a&AN=puc.210388&lang=pt-br&site=eds-live&scope=site. Acesso em: 8 set. 2024.

COOPER, R. G.; EDGETT, S. J.; KLEINSCHMIDT, E. J. Portfolio management for new products. 2. **ed. New York: Basic Books**, 2001.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv. **arXiv preprint arXiv:1810.04805**, 2019.

DONG, Li et al. Unified language model pre-training for natural language understanding and generation. **Advances in neural information processing systems**, v. 32, 2019.

DZIALLAS, Marisa; BLIND, Knut. Innovation indicators throughout the innovation process: An extensive literature analysis. Technovation, v. 80, p. 3-29, 2019.

FRESE, Michael; TENG, Eric; WIJNEN, Cees J. D.. Helping to improve suggestion systems: Predictors of making suggestions in companies. **Journal Of Organizational Behavior.** [S.l], p. 1139-1155. dez. 1999.

GABRIEL, A. MONTICOLO; D., CAMARGO, M.; BOURGAULT, M. Ontology to Represent the Knowledge Domain of a Creative Workshop. In: 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE, p. 618-623, 2016

GAMA, Fábio; FRISHAMMAR, Johan; PARIDA, Vinit. Idea generation and open innovation in SMEs: When does market‐based collaboration pay off most?. Creativity and Innovation Management, v. 28, n. 1, p. 113-123, 2019.

GERHARDT, T. E; SILVEIRA, D. T. Métodos de pesquisa. Porto Alegre: Editora da UFRGS, 2009.

GERLACH, Sophia; BREM, Alexander. Idea management revisited: A review of the literature and guide for implementation. **International Journal Of Innovation Studies.** Cologne, p. 144-161. 27 jul. 2017. Disponível em: https://www.sciencedirect.com/science/article/pii/S2096248717300048 . Acesso em: 01 jun. 2024.

GIBSON, R. SKARZYNSKI, P. Inovação: prioridade nº 1: o caminho para transformação nas organizações. Rio de Janeiro: Elsevier, 2008.

GIL, Cristina. *Clustering*. Disponível em: https://rpubs.com/cristina_gil/clustering. Acesso em: 3 dez. 2024.

GOCHERMANN, Josef; NEE, Ingo. The Idea Maturity Model—A Dynamic Approach to Evaluate Idea Maturity. International Journal of Innovation and Technology Management, v. 16, n. 05, p. 1950030, 2019.

GOOGLE RESEARCH. *Open-sourcing BERT: state-of-the-art pre-training for natural language processing*. Disponível em: https://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/. Acesso em: 3 dez. 2024.

GREEN, Stephen G.; BEAN, Alden S.; SNAVELY, B. Kay. Idea management in R&D as a human information processing analog. **Human Systems Management**, v. 4, n. 2, p. 98-112, 1983.

GURTEEN, David. Knowledge, Creativity and Innovation. **Journal Of Knowledge Management.** [S.l], p. 5-13. 1 jun. 1998. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/13673279810800744/full/htm . Acesso em: 20 abr. 2024.

HANSEN, Morten T; BIRKINSHAW, Julian. The innovation value chain. **Harvard Business Review.** [S.l], p. 121-130. jun. 2007. Disponível em: https://www.scopus.com/record/display.uri?eid=2-s2.0-34249787356&origin=inward. Acesso em: 12 abr. 2024.

HIRSCHBERG, Julia; MANNING, Christopher D. Advances in natural language processing. **Science**, v. 349, n. 6245, p. 261-266, 2015.

HOORNAERT, S.; BALLINGS, M.; MALTHOUSE, E. C.; VAN DEN POEL, D. Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time. **Journal of Product Innovation Managemen**t, v. 34, n. 5, p. 580-597, 2017.

HU, Suya; XU, Di. Identifying high quality ideas in the online context: evidence from a meta-analysis. **European Journal of Innovation Management**, v. 26, n. 3, p. 707-729, 2023.

INCITTI, F; URLI, F; SNIDARO, L. Beyond word embeddings: A survey. **Information Fusion**, v. 89, 2023.

JENSEN, Anna Vagn. A literature review of idea management. In: **DS 71: Proceedings of NordDesign 2012, the 9th NordDesign conference, Aarlborg University, Denmark. 22-24.08. 2012**. 2012.

KABIR, Muhammad Ashad et al. User-centric social context information management: an ontology-based approach and platform. **Personal and Ubiquitous Computing**, v. 18, p. 1061-1083, 2014.

KABIR, Nowshade. **The impact of semantic knowledge management system on firms' innovation and competitiveness**. 2017. Tese de Doutorado. Newcastle University.

KARIMI-MAJD, Amir-Mohsen; MAHOOTCHI, Masoud. A new data mining methodology for generating new service ideas. **Information Systems And E-Business Management.** [S.l], p. 421-443. 14 out. 2014. Disponível em: https://link.springer.com/article/10.1007/s10257-014-0267-y . Acesso em: 17 abr. 2024.

KOEN, P. A.; AJAMIAN G. M.; BOYCE, S.; CLAMEN, A.; FISHER, E.; FOUNTOULAKIS, S.; SEIBERT, R. Fuzzy front end effective methods, tools, and techniques. In: BELLIVEAU; A. P.; GRIFFIN; S. SOMERMEYER (Eds); The PDMA toolbook for new product development. New York: John Wiley, 2002.

KORNISH, Laura J.; ULRICH, Karl T. The importance of the raw idea in innovation: Testing the sow's ear hypothesis. Journal of Marketing Research, v. 51, n. 1, p. 14-26, 2014.

KOZLOWSKI, A. C; TADDY, M; EVANS, J. A. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. **American Sociological Review**, v. 84, 2019.

KULKARNI, Akshay; SHIVANANDA, Adarsha. **Natural language processing recipes**. Apress, 2019.

LAWSON, Benn; SAMSON, Danny. DEVELOPING INNOVATION CAPABILITY IN ORGANIZATIONS: A DYNAMIC CAPABILITIES APPROACH. **International Journal Of Innovation Management.** [S.l], p. 377-400. set. 2001.

LE MOIGNE, J. -L. (1995). Le constructivisme tome 2: Des épistémologies. Paris: ESF Editeur.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, v. 521, n. 7553, p. 436-444, 2015.

LEE, Tae-Young. A study on extracting ideas from documents and webpages in the field of idea mining. Journal of the Korean Society for Information Management, v. 29, n. 1, p. 25- 43, 2012.

LIDDY, Elizabeth D. Natural language processing. 2001.

LIMA, T. C. S; MIOTO, R. C. T. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica. Rev. Katál. Florianópolis. v. 10, p. 37-45, 2007.

LIN, Baihan. Knowledge management system with nlp-assisted annotations: A brief survey and outlook. **arXiv preprint arXiv:2206.07304**, 2022.

LEKA, Serena. The role of artificial intelligence in idea management systems and innovation processes: An integrative review. In: Proceedings of the Cognitive Models and Artificial Intelligence Conference. 2024. p. 160-164.

MANNING, Christopher; SOCHER, Richard. Natural language processing with deep learning. **Lecture Notes Stanford University School of Engineering**, 2017.

MARCH, Salvatore T.; STOREY, Veda C.. Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research. **Mis Quarterly.** Minnesota, p. 725-730. dez. 2008.

MARTINEZ-TORRES, Rocio; OLMEDILLA, Maria. Identification of innovation solvers in open innovation communities using swarm intelligence. **Technological Forecasting and Social Chang**e, v. 109, p. 15-24, 2016.

MCKEOWN, Max. The truth about innovation. Pearson Education India, 2008

MICHALKO, Michael. From bright ideas to right ideas: capturing the creative spark.. **The Futurist.** Washington, p. 52-56. out. 2003. Disponível em: https://www.proquest.com/openview/1d955e0a4c59a0b9d7fd00a9f59d35e5/1?pq-origsite=gscholar&cbl=47758 . Acesso em: 17 abr. 2024.

MIKELSONE, Elina et al. Idea management system application type impact on idea quantity. European Integration Studies, n. 14, p. 192-206, 2020.

MIKELSONE, Elina et al. Idea Management Type, Competencies and Capacity Impact on Innovation Results and Financial Performance. In: Harnessing AI, Machine Learning, and IoT for Intelligent Business: Volume 2. Cham: Springer Nature Switzerland, 2024. p. 3-27.

MIKELSONE, Elina; LIELA, Elita. Idea management and organisational effectiveness: A research gap. **Journal of Business Management**, v. 12, 2017.

MIKELSONE, Elina; UVAROVA, Inga; SEGERS, Jean-Pierre. Four-step approach to idea management sequencing: Redefining or reinventing values in a business model. **Journal of Innovation and Entrepreneurship**, v. 11, n. 1, p. 49, 2022.

MIKOLOV, Tomas. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MITCHELL, Melanie; KRAKAUER, David C. The debate over understanding in AI's large language models. **Proceedings of the National Academy of Sciences**, v. 120, n. 13, p. e2215907120, 2023.

MITCHELL, Tom M.; MITCHELL, Tom M. **Machine learning**. New York: McGraw-hill, 1997.

NAVEED, Humza *et al*. A Comprehensive Overview of Large Language Models. **Preprint**, [S.L.], v. [], n. [], p. 1-46, 1 jul. 2023. ArXiv. http://dx.doi.org/10.48550/ARXIV.2307.06435.

NEAGOE, Lavinia Nicoleta; KLEIN, Vladimir Mărăscu. Employee Suggestion System (Kaizen Teian) The Bottom-Up Approach For Productivity Improvement. In: **International Conference On Economic Engineering And Manufacturing Systems**. 2009, Braşov. **[S.l].** Braşov: Control, 2009. v. 10, p. 361-366.

PAULETIC, Iva; PRSKALO, Lucia Nacinovic; BAKARIC, Marija Brkic. An overview of clustering models with an application to document clustering. In: **2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. IEEE, 2019. p. 1659-1664.

PEFFERS, K; TUUNANEN, T; ROTHENBERGER, M. A; CHATTERJEE, S. A Design Science Research Methodology For Information Systems Research. **Journal of management information systems : JMIS**, v. 24, n. 3, p. 45–77, 2007.

PENNINGTON, J; SOCHER, R; MANNING C. D. GloVe: Global Vectors for Word Representation. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. pages 1532–1543, 2014.

PETUKHOVA, Alina; MATOS-CARVALHO, Joao P.; FACHADA, Nuno. Text clustering with LLM embeddings. **arXiv preprint arXiv:2403.15112**, 2024.

PINECONE. *Vector embeddings*. Disponível em: https://www.pinecone.io/learn/vector-embeddings/. Acesso em: 3 dez. 2024.

PINSONNEAULT, Alain; BARKI, Henri; GALLUPE, R. Brent; HOPPEN, Norberto. Electronic Brainstorming: the illusion of productivity. **Information Systems**

Research, [S.L.], v. 10, n. 2, p. 110-133, jun. 1999. Institute for Operations Research and the Management Sciences (INFORMS). http://dx.doi.org/10.1287/isre.10.2.110.

POVEDA, G.; WESTERSKI, A.; IGLESIAS, C. A. Application of semantic search in Idea Management Systems. International Conference for Internet Technology And Secured Transactions, 2012, vol., no., p.230 - 236, 10-12 Dec. 2012.

RAFFEL, Colin; SHAZEER, Noam; ROBERTS, Adam; LEE, Katherine; NARANG, Sharan; MATENA, Michael; ZHOU, Yanqi; LI, Wei; LIU, Peter J.. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal Of Machine Learning Research.** [S.l], p. 1-67. maio 2020. Disponível em: https://www.jmlr.org/papers/v21/20-074.html . Acesso em: 02 maio 2024.

RAMPA, Romain; AGOGUÉ, Marine. Developing radical innovation capabilities: Exploring the effects of training employees for creativity and innovation. Creativity and Innovation Management, v. 30, n. 1, p. 211-227, 2021.

RAMPA, Romain; AGOGUÉ, Marine. Developing radical innovation capabilities: Exploring the effects of training employees for creativity and innovation. Creativity and Innovation Management, v. 30, n. 1, p. 211-227, 2021.

SALDIVAR, Jorge; DANIEL, Florian; CERNUZZI, Luca; CASATI, Fabio. Online Idea Management for Civic Engagement: A Study on the Benefits of Integration with Social Networking. **Acm Transactions On Social Computing.** [S.l], p. 1-29. 23 jan. 2019. Disponível em: https://doi.org/10.1145/3284982 . Acesso em: 03 maio 2024.

SALTON, Gerard. Introduction to modern information retrieval. **McGrawHill Book Co**, 1983.

SANDSTROM, Christian; BJORK, Jennie. Idea management systems for a changing innovation landscape. **International Journal Of Product Development (Ijpd).** [S.l], p. 310-324. 05 jul. 2010.

SHINDE, Pramila P.; SHAH, Seema. A review of machine learning and deep learning applications. In: **2018 Fourth international conference on computing communication control and automation (ICCUBEA)**. IEEE, 2018. p. 1-6.

SINT, R.; MARKUS, M.; SCHAERT, S.; KURZ, T. Ideator - a collaborative enterprise idea management tool powered by KiWi. Fifth Workshop Semantic Wikis. Linking Data and People. Hersonissos, Greece, 2010.

SMILKOV, D; THORAT, N.; NICHOLSON, C.; REIF, E.; VIÉGAS, F. B.; WATTENBERG, M.. Embedding Projector: Interactive Visualization and Interpretation of Embeddings. 2016.

STENMARK, Dick. Company-wide brainstorming: Next generation suggestion systems? **Proceedings Of Iris**, [*s. l*], v. 23, n. [], p. 387-395, 12 ago. 2000.

THOM, Norberto. **Grundlagen des betrieblichen Innovationsmanagements**. 2. ed. Königstein: Peter Hanstein, 1980. 126 p. Disponível em: https://boris.unibe.ch/id/eprint/101064 . Acesso em: 10 abr. 2024.

THORLEUCHTER, Dirk; VAN DEN POEL, Dirk. Identification of interdisciplinary ideas. Information Processing & Management, v. 52, n. 6, p. 1074-1085, 2016.

TIDD, J.; BESSANT, J.; PAVITT, K. Gestão da inovação. 3ª ed. Porto Alegre: Bookman, 2005.

TURRELL, Mark. Idea management and the suggestion box. **White Paper-0802-1© Imaginatik,[online], Available www. imaginatik. Com**, 2002.

VAISHNAVi, V., & Kuechler, W. (2011). Design research in information systems. Acesso em: Maio 29, 2024. Disponível em: http://desrist.org/design-research-in-information-systems

VAISHNAVI, Vijay K. JR., William Kuechler. **Design Science Research Methods and Patterns:** Innovating Information and Communication Technology. Boca Raton: Auerbach Publications, 2008. 226 p.

VIEIRA, J. A; LEITE, A. R; KUHN, A. S. Perspectivas da Produção de Pesquisa Aplicada, Inovação e Desenvolvimento Científico e Tecnológico nos Institutos Federais. Revista Valore. v. 8, 2023.

VIERULA, Markku. What is a Competitive Advantage?. In: Find Your Market-Oriented Competitive Advantage: A Toolkit for Strategy and Branding. Cham: Springer Nature Switzerland, 2024. p. 19-28.

Wang, A.; Pruksachatkun, Y.; Nangia,N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. “Superglue: A stickier benchmark for general-purpose language understanding systems,” Advances in neural information processing systems, vol. 32, 2019.

WESTERSKI, Adam; DALAMAGAS, Theodore; IGLESIAS, Carlos A.. Classifying and comparing community innovation in Idea Management Systems. **Decision Support Systems**, [S.L.], v. 54, n. 3, p. 1316-1326, fev. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.dss.2012.12.004.

WESTERSKI, Adam; IGLESIAS, Carlos A.. Exploiting Structured Linked Data in Enterprise Knowledge Management Systems: an idea management case study. **2011 Ieee 15Th International Enterprise Distributed Object Computing Conference Workshops**, [S.L.], p. 395-403, 29 ago. 2011. IEEE. http://dx.doi.org/10.1109/edocw.2011.14.

WESTERSKI, Adam; IGLESIAS, Carlos A.; NAGLE, Tadhg. The road from community ideas to organizational innovation: a life cycle survey of idea management systems. **International Journal Of Web Based Communities (Ijwbc).** [S.I], p. 493-506. out. 2011. Disponível em: https://www.inderscienceonline.com/doi/abs/10.1504/IJWBC.2011.042993. Acesso em: 17 abr. 2024.

WILSON, Margaret A.; KREWSKI, Daniel; TYSHENKO, Michael G.. Bovine spongiform encephalopathy and variant Creutzfeldt-Jakob disease risk management in Switzerland. **International Journal Of Risk Assessment And Management.** [S.I], p. 212-224. 18 set. 2010.

XIE, Luning; ZHANG, Pengzhu. Idea Management System for Team Creation. **Journal Of Software.** [S.I], p. 1187-1194. nov. 2010.

XUE, Linting; CONSTANT, Noah; ROBERTS, Adam; KALE, Mihir; AL-RFOU, Rami; SIDDHANT, Aditya; BARUA, Aditya; RAFFEL, Colin. MT5: a massively multilingual

pre-trained text-to-text transformer. **Arxiv**, [S.L.], 22 out. 2020. ArXiv. http://dx.doi.org/10.48550/ARXIV.2010.11934 .

YOUNG, Tom et al. Recent trends in deep learning based natural language processing. **ieee Computational intelligenCe magazine**, v. 13, n. 3, p. 55-75, 2018.

ZHAI, Chengxiang; MASSUNG, Sean. **Text Data Management and Analysis**: a practical introduction to information retrieval and text mining. [S.l]: Acm Books #12, 2016. 1286 p.

ZHANG, Lei; WANG, Shuai; LIU, Bing. Deep learning for sentiment analysis: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 4, p. e1253, 2018.

ZUBIAGA, Arkaitz. Natural language processing in the era of large language models. **Frontiers in Artificial Intelligence**, v. 6, p. 1350306, 2024.