UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Luiz Fernando Spillere de Souza

MODELO DE MINERAÇÃO DE IDEIAS UTILIZANDO TÉCNICAS DE ENGENHARIA DO CONHECIMENTO

Florianópolis
2021

Luiz Fernando Spillere de Souza

MODELO DE MINERAÇÃO DE IDEIAS UTILIZANDO TÉCNICAS DE ENGENHARIA DO CONHECIMENTO

Tese submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Engenharia e Gestão do Conhecimento.
Orientador: Prof. Dr. Alexandre Leopoldo Gonçalves
Coorientador: Prof. Dr. João Artur de Souza

Florianópolis
2021

Ficha de identificação da obra

Souza, Luiz Fernando Spillere de
MODELO DE MINERAÇÃO DE IDEIAS UTILIZANDO TÉCNICAS DE ENGENHARIA DO CONHECIMENTO / Luiz Fernando Spillere de Souza ; orientador, Alexandre Leopoldo Gonçalves, coorientador, João Artur de Souza, 2021.
146 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Mineração de Ideias. 3. Engenharia do Conhecimento. 4. Aprendizado de Máquina. 5. Representação do Conhecimento. I. Gonçalves, Alexandre Leopoldo. II. Souza, João Artur de . III. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Luiz Fernando Spillere de Souza

MODELO DE MINERAÇÃO DE IDEIAS UTILIZANDO TÉCNICAS DE ENGENHARIA DO CONHECIMENTO

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof.(a) Lia Caetano Bastos, Dr.(a)
Universidade Federal de Santa Catarina

Prof. João Bosco da Mota Alves, Dr.
Universidade Federal de Santa Catarina

Prof. Gustavo Medeiros de Araujo, Dr.
Universidade Federal de Santa Catarina

Prof. Roberto Tadeu Raittz, Dr.
Universidade Federal do Paraná

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

Documento assinado digitalmente
Roberto Carlos dos Santos Pacheco
Data: 25/01/2022 16:52:00-0300
CPF: 439.543.830-04
Verifique as assinaturas em https://v.ufsc.br

Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Programa

Documento assinado digitalmente
Alexandre Leopoldo Goncalves
Data: 25/01/2022 15:49:32-0300
CPF: 786.740.349-34
Verifique as assinaturas em https://v.ufsc.br

Prof. Alexandre Leopoldo Gonçalves, Dr.
Orientador

Florianópolis, 2021.

Este trabalho é dedicado aos meus filhos Igor e Kauã, à minha esposa Jane e aos meus pais Natal e Margareth.

AGRADECIMENTOS

Agradecer é a capacidade de reconhecer a importância do outro na sua vida.

Venho primeiramente agradecer a Deus pela existência, persistência e capacidade intelectual para o desenvolvimento dos estudos.

Agradeço minha família pela tolerância e incentivo, principalmente nos momentos em que precisei me abster do convívio em prol dos estudos. Meus queridos filhos Igor e Kauã pela compreensão, minha esposa Jane pela paciência e meus pais Natal e Margareth por me mostrarem desde cedo a importância da palavra estudo.

Um agradecimento muito especial ao meu amigo e orientador Professor Dr. Alexandre Leopoldo Gonçalves pela confiança depositada na execução deste trabalho, bem como pelo compartilhamento de conhecimento nesta nobre tarefa chamada orientação. Também agradeço ao Professor Dr. João Artur de Souza, pela sua coorientação neste trabalho.

Aos professores do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento que contribuíram com a execução do trabalho na atividade de ministrar as disciplinas e composição da banca examinadora.

À Universidade Federal de Santa Catarina pela oportunidade de cursar gratuitamente um curso de Pós-Graduação de qualidade e reconhecimento nacional.

E por fim agradeço a todos que direta ou indiretamente fizeram parte deste trabalho seja contribuindo ou incentivando.

"Pensamento flexível ou elástico é o que nos confere a capacidade de resolver novos problemas e superar as barreiras neurais e psicológicas que nos impedem de enxergar para além da ordem existente. Por exemplo: a capacidade de descartar ideias confortáveis e de nos acostumarmos à ambiguidade e à contradição; a capacidade de superar posturas mentais e reestruturar as perguntas que formulamos; a capacidade de abandonar nossas suposições arraigadas e nos abrirmos para novos paradigmas; a propensão a confiar tanto na imaginação quanto na lógica, e a gerar e integrar uma grande variedade de ideias; e a vontade de experimentar e saber como lidar com o erro."

(Leonard Mlodinow)

RESUMO

No cenário atual as organizações precisam cada vez mais pensar e buscar a inovação como forma de se manterem competitivas. A inovação, entendida como um processo impacta na criação ou atualização de algo já existente, por exemplo, um produto ou serviço, tem sua origem na geração de ideias. Entre as diversas áreas envolvidas neste processo, encontra-se a Mineração de Ideias, que objetiva a identificação ou extração de ideias a partir de bases de dados não estruturados em formato de texto, por meio de ferramental advindo da Engenharia do Conhecimento. Neste sentido, o objetivo deste trabalho consiste em propor um modelo baseado em Métodos e Técnicas de Engenharia do Conhecimento, bem como em conhecimento de especialistas para auxiliar na identificação e ordenamento (*ranking*) de ideias. Para atingir este objetivo foi realizada uma revisão sistemática da literatura com a finalidade de explicitar os elementos que pudessem promover suporte a mineração de ideias. Em seguida, estes elementos foram utilizados na proposição do modelo desta tese. Para a avaliação do modelo foram elaborados dois cenários de estudo compostos por conjuntos de dados representando ideias, projetos e textos comuns. Os resultados foram obtidos executando testes inicialmente no primeiro cenário de estudo e em seguida no segundo cenário de estudo. Para o segundo cenário o *ranking* final é composto por 112 textos (50% de ideias e 50% de projetos) referentes a 20% do segundo conjunto de dados utilizado como verificação do modelo. Considerando um cenário ideal, as 56 primeiras posições deveriam ser ocupadas por ideias. A aferição desta situação após a utilização do modelo produziu uma acurácia de 80%. Esta acurácia aumenta para 93% caso sejam consideradas as primeiras 40 posições. De maneira geral, considerando os resultados obtidos pode-se concluir que as informações produzidas pelo modelo podem auxiliar na tarefa de identificação e seleção de ideias, facilitando assim a tomada de decisão por parte de especialistas na definição de quais ideias possuem maior potencial de implementação e desenvolvimento.

Palavras-chave: Mineração de Ideias. Engenharia do Conhecimento. Aprendizado de Máquina. Representação do Conhecimento.

ABSTRACT

In the current scenario, associations need to increasingly think and seek innovation as a way to remain competitive. Innovation, understood as a process that impacts the creation or updating of something that already exists, for example, a product or service, has its origins in the generation of ideas. Among the various areas involved in this process is Idea Mining, which aims to identify or extract ideas from non-structured databases in text format, using tools from Knowledge Engineering. In this sense, the objective of this work is to propose a model based on Knowledge Engineering Methods and Techniques, as well as on expert knowledge to help identify and organize ideas. To achieve the objective, a systematic literature review was carried out with a modifier to explain the elements that could support the mining of ideas. Then, these elements were used in the proposition of the model of this thesis. For an evaluation of the model, two studies of studies composed of representative data sets, projects and common texts were elaborated. The results were raised tests formulated in the first study scenario and then in the second study scenario. For the second scenario, the final ranking is composed of 112 texts (50% of the ideas and 50% of the projects) referring to 20% of the second data set used as model verification. Consider an ideal scenario, from the first 56 columns investigated being occupied by ideas. The measurement of this situation after using the model produced an accuracy of 80%. This accuracy increases to 93% if the first 40 positions are considered. In general, considering the results obtained, it can be seen that the information produced by the model can help in the task of identifying and selecting ideas, thus facilitating decision-making by the specialty in defining which ideas have the greatest potential for implementation and development.

Keywords: Idea Mining. Knowledge Engineering. Machine Learning. Knowledge Representation.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT – Associação Brasileira de Normas Técnicas

DSR - Design Science Research

DSRM - Design Science Research Methodology

DT – Decision Trees

EC – Engenharia do Conhecimento

EGC – Engenharia e Gestão do Conhecimento

IA – Inteligência Artificial

KG – Knowledge Graph

KDT - Knowledge Discovery in Text

MI – Mineração de Ideias

ML – Machine Learning

MOOC - Massive Open Online Course

MTEC – Métodos e Técnicas de Engenharia do Conhecimento

NB – Naive Bayes

PPGEGC – Programa de Pós Graduação em Engenharia e Gestão do Conhecimento

PLN – Processamento de Linguagem Natural

UFSC – Universidade Federal de Santa Catarina

RF – Random Forest

SVM – Support Vector Machines

TM – Text Mining

WE – Word Embedding

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

Para permanecerem competitivas as organizações buscam a inovação constantemente, desenvolvendo novos produtos, serviços e processos (KARIMI-MAJD; MAHOOTCHI, 2015). A inovação é um processo que ocorre por meio da criação ou renovação de algo existente, partindo de estudos, observações e persistência na busca de soluções que sejam práticas e simples, sendo facilmente entendidas e aceitas pelos consumidores (DÍAZ-GARCÍA; GONZÁLEZ-MORENO; SÁEZ-MARTÍNEZ, 2015). E para que um processo de inovação possa ser eficaz é necessária a geração de ideias inovadoras. Portanto, a capacidade de uma organização crescer depende de sua competência em gerar novas ideias e explorá-las para o seu benefício no longo prazo (FLYNN *et al.*, 2003).

Todos os dias, as pessoas criam ideias que podem contribuir no desenvolvimento de novos produtos e serviços (KHAN *et al.*, 2014). E muitas destas ideias se encontram inseridas no meio de uma grande quantidade de informação textual acessível pela *web*. Esta informação pode ser uma fonte valiosa para os tomadores de decisão, visto que podem conter muitas ideias com potencial para resolver problemas que envolvam a tomada de decisão. O desafio reside em como identificar ideias interessantes a partir de grandes volumes de dados textuais.

O método mais comum reside na busca por documentos relevantes, realizada por um especialista humano, constituindo-se em uma tarefa difícil e demorada (THORLEUCHTER; VAN DEN POEL, 2012). Por outro lado, existem métodos e técnicas que podem auxiliar na identificação de ideias em fontes de informação textual. Entre estas, a descoberta de conhecimento em textos[1], também referenciada como mineração de texto, refere-se ao processo de extrair automaticamente informações úteis a partir de dados não estruturados na forma de textos (HOTHO *et al.*, 2005).

Desta forma, Thorleuchter, Van Den Poel e Prinzie (2010) introduziram a Mineração de Ideias (MI), entendida como um processo automático voltado à extração ou identificação de ideias novas e úteis a partir de conteúdo textual usando métodos e técnicas de mineração de texto e, a partir disso, a apresentação aos usuários do resultado obtido de maneira compreensível.

[1] Neste trabalho assume-se como equivalentes os conceitos de descoberta de conhecimento em texto e mineração de texto

O processo de mineração de ideias consiste em três etapas conforme proposto por Thorleuchter, Van Den Poel e Prinzie (2010). A primeira etapa se concentra na definição do problema onde o usuário da tarefa de mineração de ideias fornece informações que descrevam seu problema específico, chamada de etapa de descrição do problema. Na segunda etapa o usuário provê informações adicionais onde ele supõe a existência de ideias que, provavelmente, possam resolver o problema. Por fim, na terceira etapa, todos os padrões de texto extraídos são avaliados quanto à novidade e utilidade. Isso significa que tais padrões são comparados à descrição do problema, usando alguma medida de mineração de ideia específica.

A literatura apresenta alguns estudos que mostram a utilização prática da tarefa de mineração de ideias aplicando técnicas computacionais. São estudos baseados em classificação de ideias em organizações (PAUKKERI, 2009), métodos automáticos no domínio tecnológico (THORLEUCHTER; VAN DEN POEL, 2012; THORLEUCHTER; VAN DEN POEL; PRINZIE, 2010), no domínio social (THORLEUCHTER; HERBERZ; POEL, 2011) e no domínio interdisciplinar (THORLEUCHTER; VAN DEN POEL, 2016), voltados para a *web* (THORLEUCHTER; VAN DEN POEL, 2013) e combinados com outras técnicas (CHRISTENSEN *et al.*, 2017a; LEE; TAN, 2017a, 2017b; SÉRGIO; SOUZA; GONÇALVES, 2017). Sob o contexto do aperfeiçoamento das técnicas de mineração de ideias, alguns estudos propõem melhorias buscando ideias em artigos (ALKSHER *et al.*, 2018a, 2018b; LIU; GOULDING; BRAILSFORD, 2015) e buscando ideias na *web* (CHRISTENSEN *et al.*, 2017b; THORLEUCHTER; VAN DEN POEL, 2015; TRIPATHY *et al.*, 2012). Citam-se ainda estudos que avaliam os métodos de mineração de ideias utilizando estatística e avaliação por especialistas humanos (ALKSHER *et al.*, 2017; KLEIN; GARCIA, 2015), assim como estudos contemplando uma revisão geral dos métodos de mineração de ideias (ALKSHER *et al*., 2016; AYELE; JUELL-SKIELSE, 2021).

1.2 APRESENTAÇÃO DO PROBLEMA

Ideias resultam de atividade mental e são formuladas verbalmente para que possam ser representadas, compartilhadas e refinadas (KIM; MACDUFFIE; PIL, 2010). No ambiente competitivo atual, as atividades geradoras de ideias tornaram-se cada vez mais importantes para o sucesso das empresas. Elaborar, criar e utilizar novos conhecimentos gera possibilidades cada vez mais diversas, o que potencializa futuras ações a serem criadas

independentemente do porte ou segmento da organização (MORAIS; MARIA; OLIVEIRA, 2021).

As ideias constituem a força vital das empresas na geração de novos produtos ou serviços, novos modelos comerciais, novos processos, bem como na produção de estratégias organizacionais de mudança e manutenção da organização (VAN DEN ENDE; FREDERIKSEN; PRENCIPE, 2015). Neste contexto, surge a Mineração de Ideias voltada à extração de informações úteis a partir de dados textuais e sendo um método automático de identificação de ideias (AYELE; JUELL-SKIELSE, 2020).

A MI é um campo novo e promissor na área de pesquisa de recuperação de informações e mineração e textos. O objetivo principal é a utilização de métodos computacionais eficientes que caracterizem uma ideia, extraindo-a ou identificando-a a partir de grandes bases de dados não estruturadas na forma textual (ALKSHER *et al.*, 2017).

Constitui-se como um processo não trivial em que as ideias representam a matéria-prima. No entanto, existem desafios como a sobrecarga de informações e a descrição trivial de ideias, fazendo com que o processo de seleção e identificação de ideias potenciais se torne relevante e desafiador para as organizações (SÉRGIO; SOUZA; GONÇALVES, 2017). Taxas de sucesso de ideias lançadas no mercado voltadas ao desenvolvimento de novos produtos são baixas. Estima-se que apenas 3 ideias bem sucedidas de cada 1000 ideias sejam lançadas. Isso sugere que práticas tradicionais utilizadas no lançamento de produtos nas organizações são altamente ineficientes (IBRAHIM; GILMOUR, 2016).

Outro desafio da área de mineração de ideias é a crescente quantidade de informação, principalmente quando se considera a *web*. A *web* é uma valiosa fonte de informação onde muitas ideias podem ser encontradas nos mais diferentes domínios. No entanto, há uma grande dificuldade para identificar essas ideias dentro da grande quantidade de informações textuais (THORLEUCHTER; VAN DEN POEL; PRINZIE, 2010).

Uma consideração importante foi realizada por Li e Chen (2014) como forma de alerta aos trabalhos publicados. Na mineração de ideias os pesquisadores se concentram em como explorar algoritmos para extrair, a partir de textos, padrões implícitos, desconhecidos e potencialmente úteis, mas ignoram as estruturas de conhecimento existentes nos textos. O conhecimento extraído pela mineração de ideias pode ser chamado de “conhecimento bruto”. Esse conhecimento deve ser avaliado e posicionado em um domínio, para então derivar o conhecimento aceito pelos usuários ou organizações e ser considerado “conhecimento inteligente”.

As técnicas de mineração de ideias automatizadas são fundamentalmente limitadas. Isto decorre pelo fato dos algoritmos atuais de processamento de linguagem natural terem apenas uma compreensão superficial da linguagem natural ou se concentrarem apenas nos resultados sem estruturar e padronizar o processo em si (AYELE; JUELL-SKIELSE, 2020; KLEIN; GARCIA, 2015). Neste sentido, Klein e Garcia (2015) propõem a utilização de avaliadores para a análise de ideias candidatas com base em uma descrição clara dos critérios de seleção estabelecidos, e assim, realizarem uma filtragem adicional.

Sérgio e Gonçalves (2019) apresentaram uma visão geral sobre a mineração de ideias e propuseram um modelo que possibilita, após a análise de conteúdo textual, a visualização das ideias com maior potencial de investimento utilizando uma base de conhecimento, podendo ser representada no formato de uma ontologia. Ainda assim, os autores ressaltam que é necessária a intervenção do especialista para a avaliação do resultado do modelo, onde este determinará se uma ideia irá ou não ser implementada.

Especialistas em um domínio particular desenvolvem estratégias ou heurísticas que permitem selecionar ideias a partir da leitura de textos e aplicação de critérios de escolha. Os métodos computacionais por sua vez, realizam esta mesma tarefa de seleção de ideias, utilizando algoritmos concebidos para executar a tarefa de um especialista de domínio. Um mecanismo capaz de traduzir esse conhecimento de domínio dos especialistas humanos em uma estrutura de aprendizagem coesa e expressiva, combinando conhecimento dos seres humanos com a potencialidade das máquinas, constitui-se em uma pesquisa desafiadora (SILVA; GOMBOLAY, 2019).

Neste sentido, o presente trabalho, ao lidar com os desafios acima descritos recorre a Métodos e Técnicas de Engenharia do Conhecimento (MTECs). Mais especificamente, esta tese baseia-se nos conceitos de Aprendizado de Máquina, por meio da classificação de texto, Incorporação de Palavras (*Word Embeddings* - WE*)* e Grafos de Conhecimento (*Knowledge Graphs* - KG*),* com o intuito de integrar elementos computacionais e de conhecimento humano especializado (critérios de seleção de ideias) que possam auxiliar na mineração de ideias.

A classificação de texto é uma tarefa de aprendizagem supervisionada em que um modelo é treinado para prever rótulos para documentos desconhecidos a partir da observação de documentos rotulados (MOREO; ESULI; SEBASTIANI, 2021). Entre os classificadores citam-se as Árvores de Decisão (*Decision Trees* - DT*)*, o *Naive Bayes (*NB*)*, as Máquinas de

Vetores de Suporte (*Support Vector Machines* – SVM*)* e as Florestas Aleatórias (*Random Forest* – RF*)*.

Já os WE codificam relações de palavras dentro de um espaço vetorial (HEIMERL; GLEICHER, 2018) e têm se mostrado eficazes para capturar informações semânticas latentes em várias tarefas de processamento de linguagem natural (SHUANG *et al.*, 2020). Por sua vez, os KG representam o conhecimento humano no formato de grafo capaz de ser entendido por máquinas e estão se tornando um importante elemento para várias aplicações de inteligência artificial e processamento de linguagem natural (WANG *et al.*, 2018), incluindo recuperação de informação, resposta à perguntas e geração de hipóteses (ZHANG; HUANG; TAN, 2020).

A partir da análise de diversos trabalhos, alguns já citados acima e considerando o ponto de vista de seus autores, a seguir são relacionados os principais desafios encontrados:

- A partir de certa quantidade de informação textual torna-se difícil para o ser humano lidar adequadamente com a identificação de ideias (SÉRGIO; SOUZA; GONÇALVES, 2017; THORLEUCHTER; VAN DEN POEL; PRINZIE, 2010);
- Os pesquisadores se concentram em como explorar algoritmos para extrair, a partir de textos, padrões implícitos, desconhecidos e potencialmente úteis, mas ignoram as estruturas de conhecimento existentes nos textos (AYELE, 2020; LI; LI; CHEN, 2014);
- O conhecimento do especialista é fundamental na etapa de seleção de ideias. Existe a necessidade de traduzir esse conhecimento de domínio dos especialistas humanos em uma estrutura de aprendizagem coesa e expressiva, pois os especialistas em um domínio particular desenvolvem estratégias ou heurísticas que permitem superar os métodos computacionais na tarefa de seleção de ideias (SÉRGIO; GONÇALVES, 2019; SILVA; GOMBOLAY, 2019);
- A identificação dos critérios utilizados pelos especialistas na seleção de ideias, bem como o seu uso na tarefa de mineração de ideias através de MTECs é outro ponto relevante. Critérios adequados para selecionar ideias aumentam as chances de organizações desenvolverem produtos potencialmente inovadores e competitivos (VALDATI; DANDOLINI, 2019).
- A importância e as possibilidades de geração e avaliação de ideias estão aumentando devido às crescentes demandas por inovação digital e pelo aumento

de dados textuais disponíveis. Desta forma, a mineração de ideias assume papel importante. O desafio atual consiste em utilizar ferramentas de aprendizado de máquina e análise visual no apoio à geração e avaliação de ideias (AYELE; JUELL-SKIELSE, 2020).

1.3 PERGUNTA DE PESQUISA

Considerando o exposto acima, esta tese objetiva responder a seguinte pergunta de pesquisa: “Como identificar ideias a partir de fontes de dados não estruturados na forma de texto, considerando critérios de escolha utilizados por especialistas durante o processo de seleção de ideias?”.

1.4 OBJETIVOS

1.4.1 Objetivo Geral

O objetivo geral deste trabalho é propor um modelo voltado à identificação de ideias a partir de fontes de dados não estruturados na forma de texto levando-se em conta MTECS e critérios de escolha de ideias utilizados por especialistas.

1.4.2 Objetivos Específicos

- Identificar os principais critérios utilizados por especialistas na identificação de ideias;
- Selecionar os principais métodos e técnicas de mineração de ideias que possam promover suporte ao desenvolvimento do modelo proposto considerando os critérios utilizados por especialistas;
- Apresentar as ideias identificadas, em formato de *ranking*, em que fique explicitada a aderência da ideia aos critérios utilizados pelos especialistas;
- Evidenciar a viabilidade do modelo proposto através do desenvolvimento de um protótipo e aplicação deste em cenários de estudo.

1.5 JUSTIFICATIVA E RELEVÂNCIA DO TEMA

A capacidade organizacional para inovar constantemente, recriando e modificando recursos internos e externos é a principal fonte de sobrevivência das empresas em ambientes competitivos (NISULA; KIANTO, 2013; SILVA; PEDRON, 2019). No atual cenário econômico, as mudanças tecnológicas e de mercado são altamente imprevisíveis e contínuas. Por isso, a inovação é essencial para as organizações, agregando valor a produtos e serviços, mantendo a empresa no mercado e gerando vantagem competitiva (COSTA *et al.*, 2021).

O potencial inovador de uma organização está diretamente relacionado com o lançamento e aceitação de seus produtos ou serviços pelo mercado, que é a evidência da eficácia dos processos de inovação na organização (SADRIEV; PRATCHENKO, 2014).

A inovação tornou-se a chave para o sucesso de muitas organizações sendo pré-requisito de competitividade no mundo atual. Impulsionada pela capacidade de seus colaboradores em gerar ideias interessantes e, a partir disso, realizar a etapa de implementação (ALKSHER *et al.*, 2018a).

As organizações empresariais acumulam muito conhecimento, incluindo grandes bancos de dados sobre produtos, clientes, concorrentes, mercados, entre outros, reunidos ao longo dos anos. As organizações também têm conhecimento pessoal arquivado no espaço de trabalho do colaborador, como arquivos de texto, dados, livros e *e-mails* (NONAKA, IKUJIRO; TAKEUCHI, 2000).

E, inseridas nesses bancos de dados, que podem ou não estar conectados à *internet*, muitas vezes encontram-se boas ideias. Essas ideias, se não exploradas, perdem a sua finalidade, pois nunca serão colocadas em prática ou até mesmo corre-se o risco de perdê-las.

Entretanto, convocar um grupo de especialistas para identificar as melhores ideias, a partir de grandes bases textuais, pode ser proibitivamente caro e lento (KLEIN; GARCIA, 2015).

Com o aumento no volume desses dados passa a ser impraticável a leitura e análise somente com métodos manuais. Usando métodos de mineração de texto é possível processar grandes quantidades de dados textuais e encontrar conexões semânticas entre eles. Para realmente apoiar a criatividade e inovações nas organizações, devem existir ferramentas de *software* que coletem e armazenem ideias para então torná-las gerenciáveis (AYELE; JUELL-SKIELSE, 2020; PAUKKERI, 2009).

A mineração de ideias é uma tendência de pesquisa em engenharia e extração de conhecimento. Ela depende principalmente de informações latentes e de suas mudanças dinâmicas para conduzir à criação, integração e avaliação de ideias. O processo de mineração de ideias é baseado na extração de termos e relacionamentos latentes em dados não estruturados identificados como padrões de texto (ALKSHER *et al.*, 2017).

Outro fator a ser considerado é que muitas inovações falham porque as ideias não inovadoras são selecionadas como ponto de partida. Isso torna o processo de inovação demorado e dispendioso (DISSELKAMP, 2015). A solução adequada para promover inovação vai além de um *software*, reside na escolha correta de um modelo ou processo para o gerenciamento da inovação com uma seleção e implementação das ideias alinhadas à realidade e cultura organizacional das empresas (COSTA *et al.*, 2021).

Dentro deste contexto, o presente trabalho se justifica pela sua contribuição no aprimoramento da tarefa de mineração de ideias a partir da utilização de critérios de especialistas e de MTECs. A proposta está situada em uma pesquisa interdisciplinar interconectando a Mineração de Ideias e a Engenharia do Conhecimento.

1.6 ORIGINALIDADE

Para garantir a originalidade e a relevância deste trabalho, foi realizada uma revisão sistemática de literatura utilizando as bases de dados Scopus®, Science Direct®, ACM®, IEEE® e Springer Link®. No capítulo 3 serão apresentados os detalhes metodológicos utilizados nesta revisão. A partir desta revisão sistemática da literatura foi possível constatar que:

- Dos estudos encontrados, pode-se dizer que todos se dizem eficazes, inclusive com demonstração de estudos de caso, o que responde ao questionamento inicial desta revisão. Todavia, a mineração de ideias ainda é uma área de conhecimento que pode ser bastante explorada, tendo em vista sua complexidade. Para que um método computacional possa efetivamente identificar uma ideia, é necessário que este esteja embasado sobre o que é uma ideia e como ela pode ser identificada em um texto. Por exemplo, ao se investigar detalhadamente o processo de identificação de ideias realizado por especialistas humanos e, compreendendo quais os critérios são levados em consideração, torna-se possível aprimorar os métodos de mineração de ideias. Para se obter um método, técnica, modelo, entre

outros, que conduza a um elevado grau de acurácia é fundamental conhecer e sistematizar como o processo é realizado por especialistas.

- Outro ponto relevante é a apresentação da aderência das ideias encontradas aos critérios dos especialistas. O trabalho pretende apresentar, como resultado final, um ordenamento (*ranking*) de ideias considerando os critérios dos especialistas para determinar quais textos podem ser classificados como uma ideia. Dentro da revisão sistemática realizada, não foram encontrados trabalhos que explicitassem esta possibilidade.
- Destaca-se que nos últimos anos trabalhos que combinam mineração de ideias com a utilização de técnicas de aprendizado de máquina para treinar algoritmos capazes de identificar ideias têm sido elaborados. Este parece ser um campo promissor com espaço para aperfeiçoamentos e que possui vantagens e desvantagens em relação à técnicas mais tradicionais. Pelo fato de utilizarem aprendizado supervisionado, o conceito de ideia torna-se simplificado, visto que são apresentados ao sistema exemplos de ideias selecionadas por especialistas dentro de um conjunto de dados. Porém, essa seleção de ideias pode ser subjetiva e tornar o método pouco efetivo, uma vez que o sucesso da aprendizagem de máquina está altamente relacionado à qualidade dos dados.

Avaliando as constatações obtidas na revisão sistemática da literatura, esta tese propõe um modelo de mineração de ideias utilizando MTECs, principalmente a representação do conhecimento e o aprendizado de máquina. Além da contribuição principal, outras contribuições do modelo podem ser citadas:

- Separação das ideias de acordo com os critérios dos especialistas utilizados durante o processo de identificação e/ou seleção de ideias com potencial de implementação;
- Utilização de técnicas de representação de conhecimento e aprendizado de máquina para modelar a forma como uma ideia é escolhida por um especialista humano;
- Apresentação de um conjunto de ideias na forma de uma classificação (*ranking*), com o intuito de auxiliar especialistas na escolha de quais devem ser implementadas.

- Desenvolvimento de um protótipo que ateste a viabilidade do modelo proposto através da demonstração em cenários de estudo.

1.7 ESCOPO DO TRABALHO

O escopo deste trabalho constitui-se na proposição e desenvolvimento de um modelo capaz de agregar à tarefa de mineração de ideias diversas técnicas, objetivando contribuir no processo de identificação de ideias realizado por seres humanos.

Para o modelo proposto serão utilizados MTECs, mais especificamente, o Aprendizado de Máquina, por meio de classificadores de texto, entre eles, DT, NB, SVM e FT, bem como os WEs e KGs para promover suporte à representação de conhecimento. Estas MTECs serão utilizadas como ferramentas para identificação de ideias, de acordo com os critérios utilizados por especialistas.

A tese não pretende evoluir na questão dos critérios dos especialistas, sendo que estes serão buscados através de revisão da literatura. Esta tese também não possui o intuito de prover um produto final, mas sim um *software* na forma de protótipo para a avaliação do modelo através de cenários de estudo.

As bases de dados utilizadas serão bases de ideias disponíveis na *web*. Estas serão coletadas e armazenadas em bancos de dados para utilização pelo modelo. Os textos estarão escritos no idioma português, bem como todas os MTECs utilizados deverão considerar este idioma. Os conjuntos de dados utilizados para treinamento e testes não são considerados grandes devido à dificuldade de se encontrar bases de ideias disponíveis para consulta. Ainda assim, a análise, classificação e seleção de ideias é complexa para o ser humano representando um desafio caso fossem realizadas sem o auxílio computacional.

Vale ressaltar que a premissa desta tese é que o conhecimento de especialistas utilizado para selecionar ideias pode ser explicitado e modelado através da representação de conhecimento e aprendizado de máquina, promovendo avanços na área de Mineração de Ideias.

1.8 ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO

Esta seção tem por objetivo contextualizar a aderência do trabalho desenvolvido na tese ao objeto de pesquisa do Programa de Pós-Graduação em Engenharia e Gestão do

Conhecimento (PPGEGC/UFSC). Para torná-la mais explicativa, dividiu-se a seção em três subseções que apresentam a identidade da tese, o contexto estrutural da pesquisa no PPGEGC e as referências factuais da pesquisa a partir da base de teses e dissertações do PPGEGC.

1.8.1 Identidade

A proposta desta tese está centrada na área de Engenharia do Conhecimento (EC) voltada à tarefa de mineração de ideias com reflexo na solução de problemas apresentados pela Mineração de Ideias.

A Mineração de Ideias é vista como a aplicação de técnicas e algoritmos que buscam identificar padrões textuais relevantes em um conteúdo não estruturado para resolver um determinado problema estratégico (THORLEUCHTER; VAN DEN POEL, 2013).

A Engenharia do Conhecimento é uma área que se dedica ao desenvolvimento e uso de ferramentas que auxiliem nos processos de extração e modelagem do conhecimento, de forma a torná-lo explícito, formalizando, disponibilizando e disseminando o conhecimento em um sistema de conhecimento. A Engenharia do Conhecimento abrange métodos e técnicas de várias disciplinas, como Inteligência Artificial, Aprendizado de Máquina, Banco de Dados, Mineração de Dados, entre outras, além de utilizar métodos e técnicas que permitem a elicitação do conhecimento de especialistas a fim de modelá-los, visando a automatização de tarefas, a representação de processos e o reconhecimento de padrões (VIEIRA, 2020).

Esta tese vem ao encontro do exposto acima, pois pretende modelar conhecimento para automatizar tarefas, tendo como base a formalização do conhecimento de especialistas.

1.8.2 Contexto Estrutural da Pesquisa no PPGEGC

Além da área de concentração de Engenharia do Conhecimento, esta tese se relaciona à linha de pesquisa Teoria e Prática em Engenharia do Conhecimento, pois aborda metodologias e tecnologias da Engenharia do Conhecimento e da Inteligência Computacional e suas relações com a Gestão do Conhecimento (EGC-UFSC, 2021).

Este enquadramento ocorre devido ao estudo e aplicação de técnicas e metodologias que objetivam contribuir significativamente na infraestrutura do conhecimento de uma organização, originando subsídios que auxiliam a Gestão do Conhecimento (GONÇALVES, 2006).

O objeto de pesquisa do PPGEGC é o conhecimento que deve ser analisado sob uma perspectiva interdisciplinar. Neste sentido, o Programa articula conexões interdisciplinares destacando três ênfases de formação e pesquisa:

> (1) explicitação, emulação e modelagem do conhecimento, englobando a criação, a descoberta, a aquisição, a formalização, a codificação, o armazenamento, a distribuição e uso de conhecimento (Engenharia); (2) planejamento e alinhamento coletivo e organizacional do conhecimento compostos de sub processos como integração, avaliação, auditoria, retenção-descarte, criação-inovação, propriedade e evolução do conhecimento (Gestão); e (3) difusão, comunicação e compartilhamento do conhecimento que abrange a preservação, disseminação, transferência, socialização e acesso ao conhecimento (Mídia) (EGC-UFSC, 2021).

Dentre os vários conceitos de interdisciplinaridade, um deles é a construção de um sistema complexo que visa integrar as verdades de cada disciplina como unidades simples, mas aceitando suas diferenças e respeitando a complexidade de sua própria formação, reintegrando cada disciplina em um todo que já foi um dia naturalmente unido (PACHECO; TOSTA; FREIRE, 2010).

A aderência deste trabalho ao objeto de pesquisa do PPGEGC, ou seja o conhecimento, pode ser reforçada a partir da área de concentração de Engenharia do Conhecimento:

> A EC define conhecimento como “processo e produto tangível ou intangível efetivado na relação entre pessoas e agentes não humanos para a geração de valor.” Desta forma, com base na visão cognitivista, os objetivos desta área de concentração incluem a pesquisa e o desenvolvimento de métodos, técnicas e ferramentas para a construção de modelos e sistemas de conhecimento em atividades intensivas em conhecimento (EGC-UFSC, 2021).

Conforme o exposto, a presente tese está aderente ao objeto do programa, o conhecimento, pois propõe um modelo de mineração de ideias utilizando-se de MTEC. O enfoque da pesquisa é interdisciplinar, pois busca conceitos teóricos das disciplinas Engenharia do Conhecimento, Ciência da Computação em colaboração com a Mineração de ideias.

1.8.3 Referências Factuais

O Programa PPGEGC possui trabalhos desenvolvidos que são relacionados com a área de Mineração de Ideias, sendo estes:

- ALVAREZ, Guilherme Martins. Análise de Agrupamentos e Mineração de Opinião como Suporte à Gestão de Ideias. Dissertação (Mestrado) - Programa de

Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2018.

- DOROW, Patrícia Fernanda. O Processo de Geração de Ideias para Inovação: Estudo de Caso em uma Empresa Náutica. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2013.
- MIGUEZ, Viviane Brandão. Uma Abordagem de Geração de Ideias para o Processo de Inovação. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2012.
- PRADA, Charles A. Proposta de modelo para o gerenciamento de portfólio de inovação: modelagem do conhecimento na geração de ideias. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2009.
- RIBEIRO, Alessandro Costa. Modelo de Reconhecimento de Padrões em Ideias usando Técnicas de Descoberta de Conhecimento em Textos. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.
- SÉRGIO, Marina Carradore. Um Modelo Baseado em Ontologia e Análise de Agrupamento para Suporte à Gestão de Ideias. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.
- SÉRGIO, Marina Carradore. Modelo de Avaliação de Potenciais Ideias Alinhadas ao Contexto Organizacional. Tese (Doutorado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2020.

No que tange ao estudo do processo de critérios de avaliação de ideias, o programa PPGEGC também possui trabalhos desenvolvidos, sendo estes:

- ROCHADEL, Willian. Identificação de Critérios para Avaliação de Ideias: Um Método Utilizando Folksonomias. Dissertação (Mestrado) - Programa de Pós-

graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.

- VALDATI, Aline de Brittos. Processo de seleção de ideias em empresas inovadoras. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2017.

E por último, cabe citar os trabalhos desenvolvidos no PPGEGC na área de Descoberta de Conhecimento:

- BOVO, Alessandro Botelho. Um Modelo de Descoberta de Conhecimento Inerente à Evolução Temporal dos Relacionamentos Entre Elementos Textuais. Tese (Doutorado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2011.
- WOSZEZENKI, Cristiane Raquel. Modelo para Descoberta de Conhecimento Baseado em Associação Semântica e Temporal entre Elementos Textuais. Tese (Doutorado) - Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2016.

A partir das referências apresentadas, a presente tese é aderente ao Programa EGC por evidenciar que o trabalho está de acordo com a área de concentração de Engenharia do Conhecimento, promovendo suporte à área de Mineração de Ideias, e possui trabalhos anteriores que abordam temáticas similares dentro do mesmo programa.

1.9 ORGANIZAÇÃO

O trabalho é composto por seis capítulos que descrevem todo o percurso da tese.

No capítulo 1 é apresentada uma introdução ao tema para situar o leitor a respeito do problema de pesquisa, da pergunta de pesquisa, dos objetivos, da justificativa, da originalidade, da delimitação e da aderência ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC/UFSC).

O referencial teórico é apresentado no capítulo 2 e foi obtido através de uma visão interdisciplinar, objetivando estudar os métodos e as técnicas da Engenharia do Conhecimento e suas relações com a Mineração de Ideias, auxiliando na tomada de decisão no processo de identificação e seleção de ideias. Os instrumentos empregados para o referencial teórico

foram a pesquisa bibliográfica com foco na revisão sistemática, utilizando bases de dados internacionais como Scopus®, Web of Science®, ACM®, IEEE® e Springer Link®.

O capítulo 3 detalha o enquadramento metodológico, a aplicação da metodologia *Design Science Research* e a revisão sistemática realizada. No capítulo 4 apresenta-se o modelo proposto e a descrição detalhada das etapas que o compõem.

No capítulo 5 é apresentada como foi realizada a avaliação do modelo seguido da discussão dos resultados alcançados por meio de cenários de estudo.

E por fim, o capítulo 6 explicita as conclusões da tese, assim como as sugestões identificadas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os tópicos que fundamentam a pesquisa elaborada, sendo eles: Mineração de Texto, Mineração de Ideias, Aprendizagem de Máquina, Classificação de Texto, *Word Embeddings*, *Knowledge Graph* e Trabalhos Correlatos na área de Mineração de Ideias. Ao final deste capítulo são apresentadas as considerações finais. A seguir será apresentado o referencial teórico que envolve cada um dos tópicos.

2.1 MINERAÇÃO DE TEXTO

A mineração de texto (do inglês *Text Mining* – TM) ou descoberta de conhecimento a partir de textos (do inglês *Knowledge Discovery in Text* – KDT) refere-se ao processo de extrair automaticamente informações interessantes e conhecimento de conteúdo não estruturado (HOTHO *et al.*, 2005).

O conceito de mineração de texto foi introduzido pela primeira vez por Feldman e Dagan (1995) e abrange uma grande quantidade de métodos, técnicas e algoritmos para análise de texto, aplicados em diferentes campos, onde cita-se recuperação de informação, processamento de linguagem natural, mineração de dados e aprendizado de máquina (ALLAHYARI *et al.*, 2017).

Em muitas aplicações de mineração de texto, particularmente na recuperação de informação, os documentos precisam ser classificados de forma a definir a importância de uma palavra em um documento. A forma mais comum de fazer esta definição desta importância é representar os documentos como vetores e um valor numérico é atribuído a cada palavra do documento, de forma a graduar esta importância (SINGHAL, 2001).

As aplicações da mineração de texto são abrangentes e impactam em vários domínios, estando presentes em uma ampla gama de áreas de pesquisa, com o propósito de organizar e compreender documentos, revelando padrões semânticos em um texto (KARAMI *et al.*, 2020).

A mineração de texto tem sido usada em vários estudos que envolvem extração de informação para tomada de decisão (PEJIC-BACH *et al.*, 2020), dentre os quais pode-se citar a mineração de ideias, seção tratada a seguir.

2.2 MINERAÇÃO DE IDEIAS

Com base no conceito e mineração de texto, Thorleuchter, Van Den Poel e Prinzie (2010) introduziram o conceito de mineração de ideias como um processo automático voltado à extração e seleção de ideias novas e úteis de texto não estruturado usando métodos e técnicas de mineração de texto e, a partir disso, apresentar ao usuário o resultado obtido de maneira compreensível.

O processo de mineração de ideias inicialmente proposto por Dirk Thorleuchter *et al.* (2010) pode ser descrito em 3 etapas. A primeira etapa se concentra no problema, onde o usuário tem que fornecer informações textuais onde ele descreve seu problema específico. Na segunda etapa o usuário deve fornecer mais informações textuais que provavelmente, possam resolver o seu problema. Finalmente, a terceira etapa consiste em avaliar todos os padrões de texto extraídos quanto à novidade e utilidade. Isso significa que eles são comparados à descrição do problema usando uma medida de mineração de ideia específica capaz de classificar os textos retornados como uma ideia nova e útil.

Pouco depois, Thorleuchter e Van Den Poel (2012) evoluíram o conceito anteriormente descrito de mineração de ideias como sendo o processo em que se divide uma ideia em um meio e um fim, através da criação de um padrão textual representativo para o meio e o fim, e que identifica novas ideias procurando por meios desconhecidos que aparecem juntamente com fins conhecidos ou vice-versa. Meios e fins que ocorrem no mesmo contexto de uma descrição do problema são meios e fins conhecidos e representam uma ideia conhecida. Caso contrário, se meios e fins puderem ser encontrados e não ocorrerem neste mesmo contexto de uma descrição do problema, então eles são desconhecidos e essa combinação representa uma nova ideia (THORLEUCHTER; VAN DEN POEL, 2013).

Está técnica ainda é utilizada como base para o surgimento de melhorias no processo. Trabalhos como Alksher *et al.* (2018b) consideram as relações semânticas entre os termos conhecidos e desconhecidos dentro do padrão de texto, melhorando o desempenho da tarefa de mineração de ideias.

E num contexto ainda mais atual, a de mineração de ideias evoluiu para o uso de técnicas mais específicas, como técnicas de aprendizado de máquina, aprendizagem profunda, recuperação de informações, filtragem colaborativa, regras de associação, etc., para gerar ideias úteis e novas a partir de dados textuais não estruturados ou semiestruturados (AYELE, 2020).

2.2.1 Definição de Ideia

As organizações que desejam manter a inovação constante precisam de um fluxo contínuo de novas ideias. Essa necessidade impulsionou o uso de plataformas com o objetivo de gerar novas ideias e incentivar os funcionários e clientes a participarem do processo de inovação (ELERUD-TRYDE; HOOGE, 2014).

Todos os dias, as pessoas criam ideias que podem contribuir no desenvolvimento de novos produtos e serviços (KHAN *et al.*, 2014). Muitas destas ideias se encontram inseridas no meio de uma grande quantidade de informação textual acessível pela *internet*. Esta informação pode ser uma fonte valiosa para os tomadores de decisão, visto que podem conter muitas ideias interessantes com potencial para resolver problemas que envolvam a tomada de decisão. O desafio reside em como extrair ideias interessantes a partir de grandes volumes de informação textual. O método mais comum é a extração intuitiva por um especialista humano, constituindo-se em uma tarefa difícil e demorada (THORLEUCHTER; VAN DEN POEL, 2012). Por outro lado, existem métodos e técnicas capazes de auxiliar na identificação de ideias em fontes de informação textual.

Uma ideia é uma imagem existente ou formada na mente, mas que pode ser escrita como informação textual (THORLEUCHTER; VAN DEN POEL; PRINZIE, 2010). Outra definição, mais filosófica, define ideia como uma combinação de duas coisas: um significado e um propósito específico (POZZO, 1998). Já Azman *et al.* (2020) define uma ideia como sendo a combinação de um meio desconhecido e um propósito conhecido. E que tais condições dentro de um texto aumentam a probabilidade deste conter uma nova ideia para o leitor.

2.2.2 A Mineração de Ideias como Método de Extração e Seleção de Ideias

As organizações empresariais acumulam muito conhecimento, incluindo grandes bancos de dados sobre produtos, clientes, concorrentes, mercados, entre outros, reunidos ao longo dos anos. As organizações também têm conhecimento arquivado no espaço de trabalho pessoal de seus colaboradores, como arquivos de texto, dados, livros e *e-mails* (NONAKA, IKUJIRO; TAKEUCHI, 2000; TREVISAN; PELOGIA; DAMIAN, 2018).

A tendência é que esses bancos de dados cresçam rapidamente em volume e, desta forma, tornam-se impraticáveis a realização de análises manuais. Usando métodos avançados

de mineração de texto é possível processar volumosas quantidade de dados e encontrar conexões semânticas entre eles. Para realmente apoiar a criatividade e inovações nas organizações existem ferramentas de *software* que coletam e armazenem ideias para então torná-las gerenciáveis e úteis em um processo de tomada de decisão (PAUKKERI, 2009).

A mineração de ideias é uma área de pesquisa que depende principalmente de informações latentes e suas mudanças dinâmicas para conduzir a criação de ideias, integração e avaliação durante o processo de criatividade sustentável. O processo de mineração de ideias é baseado na extração de termos latentes no texto a partir de texto não estruturado, utilizando padrões de caracterização da ideia (ALKSHER *et al.*, 2017).

Devido ao volume de dados textuais, é impossível resumir as informações manualmente e, consequentemente, métodos automáticos eficientes de mineração de texto são necessários para se extrair ideias (ALKSHER *et al.*, 2016).

A avaliação humana é frequentemente necessária no processamento de linguagem natural para avaliar desempenho do sistema (KISHIDA, 2005), sendo importante considerar a seleção dos avaliadores, em termos de especialização e experiência (ALKSHER *et al.*, 2017). Em um contexto ideal, um método automático de mineração de ideias deve incorporar critérios de seleção de ideias utilizados por especialistas, para que o sistema possa funcionar da forma mais semelhante possível com as escolhas desses especialistas. Estes critérios podem representar fatores quantitativos e qualitativos a serem considerados, por exemplo: disponibilidade de recursos, necessidade do mercado, superioridade do produto, singularidade, complexidade tecnológica e riscos sobre resultado de projetos (MOUSAVI; TORABI; TAVAKKOLI-MOGHADDAM, 2013).

2.2.3 Critérios de Seleção de Ideias

A atividade de seleção de ideias pode ser realizada por especialistas em uma organização. Segundo Cooper e Edgett (2008), um grupo de gerentes pode se reunir periodicamente para analisar as ideias e avaliá-las por meio de um sistema de pontuação, que determina critérios preestabelecidos para essa decisão. Se a ideia for rejeitada, o criador da ideia recebe um *feedback* justificando, a partir dos critérios, o porquê da não aceitação. Isso permite ao autor reformular a ideia para participar novamente do processo, garantindo um fluxo constante de novas ideias.

Magnusson, Wästlund e Netz (2014) resumem esta corrente de pesquisa na não existência uniforme de critérios aceitos para a seleção de ideias. Assim, os critérios devem ser escolhidos em função do contexto e da fase do ciclo de desenvolvimento.

Ferioli *et al.* (2008) tentaram estimar uma lista não exaustiva que apresenta um conjunto de critérios considerados como os mais importantes por sua investigação. Os autores ainda afirmam que os critérios sejam bem definidos para permitirem que os especialistas os usem no momento da seleção de forma eficiente.

Quando não existiam ferramentas formais de avaliação de ideias, os especialistas usavam sua sensibilidade e experiência para avaliar as ideias geradas pela criatividade, onde então era possível que uma boa ideia escapasse da sua sensibilidade (FERIOLI *et al.*, 2008).

Existem diversos trabalhos publicados que tratam de estabelecer critérios de extração de ideias (FERIOLI *et al.*, 2008; MAGNUSSON; WÄSTLUND; NETZ, 2014; VALDATI, 2017). Não é o foco desta tese estabelecer ou criar critérios, isto será feito através de revisão de literatura, onde serão estabelecidos alguns critérios de uso, por ordem de importância dentro do aspecto tecnológico.

Diante do exposto acima, constata-se que os critérios mais recorrentes na literatura são: Viabilidade (Produtibilidade, Compatibilidade de Recursos e Compatibilidade com a Infraestrutura da Empresa) e Originalidade (subdividida em Novidade, Criatividade e Singularidade do Produto). Estes critérios fazem parte do aspecto tecnológico, possuindo o maior número de critérios na literatura pesquisada, seguido do aspecto Econômico. Este aspecto engloba questões de mercado financeiro e clientes. Com menor frequência são citados os aspectos subjetivo e social (VALDATI, 2017).

A partir desta análise da literatura foram identificados os três critérios mais recorrentes na literatura, listados e brevemente definidos abaixo:

- Produtibilidade: representa a perspectiva da empresa sobre a facilidade com que o serviço pode ser implementado ou produzido. Este critério assume a perspectiva da oferta (MAGNUSSON; WÄSTLUND; NETZ, 2014).
- Originalidade: representa a novidade e inovação, isto é, o quão incomuns e originais as ideias são a respeito do contexto em que se inserem (MAGNUSSON; WÄSTLUND; NETZ, 2014). O autor baseia-se em Amabile (1996), no qual defende que a originalidade é um conceito genérico e que as pessoas têm uma sensação intuitiva para o que é criativo.

- Viabilidade Econômica: Para aprofundar o carácter econômico durante a avaliação da ideia, é essencial estimar o preço de venda e o benefício potencial. Também é importante avaliar se está de acordo com os objetivos da empresa e estimar os custos de produção e de desenvolvimento, bem como o tempo que vai levar para atingir o mercado (FERIOLI *et al.*, 2008; OZER, 2004).

A pesquisas na área de seleção de ideias mostraram que não existe um critério simples para esta seleção, revelando que um conjunto complexo de critérios diferentes devem ser considerados. Critérios estes que são interdependentes e que influenciam fortemente um ao outro. No entanto, ao criar uma estrutura de suporte para seleção de ideias, deve-se considerar cada critério individualmente para promover objetividade à estrutura (GRÜNLING, 2017).

A determinação dos critérios, bem como sua seleção, normalmente é realizada subjetivamente pelos tomadores de decisão de uma organização, onde uma avaliação final dos critérios geralmente é realizada por um grupo de avaliadores (OLSSON; LANDSTRÖM, 2020).

As soluções automatizadas podem ser úteis no processo de seleção de ideias. Como por exemplo, a automatização da pré-seleção de ideias, em que se as ideias que não atenderem determinados pré-requisitos são eliminadas. Como esta tarefa de seleção inicial exige muito esforço manual, o desenvolvimento tecnológico de áreas como inteligência artificial, processamento de linguagem natural e aprendizado de máquina, atuam como ferramentas para desenvolver serviços automatizados que podem ajudar indivíduos durante o processo seleção de ideias (SARIGIANNI *et al.*, 2020).

Como exemplo bem-sucedido pode-se citar o trabalho de Li *et al*. (2021) que desenvolveu métodos automáticos de seleção de ideias para novos produtos focado na abordagem de tomada de decisão multicritério, na qual critérios múltiplos e subcritérios podem ser considerados no processo de seleção. As abordagens utilizadas para a tomada de decisão multicritério foram a teoria dos conjuntos difusos e o método do melhor-pior (*Best-Worst Method*). O método do melhor-pior (REZAEI, 2015) é usado para avaliar um conjunto de alternativas com relação a um conjunto de critérios de decisão.

Diante do exposto acima e, considerando que a identificação de critérios de seleção de ideias é subjetiva, complexa e interdependente e, que a automatização do processo de seleção de ideias mostra ser viável, a proposta da tese no que tange a identificação de critérios

consiste na utilização de MTECs, envolvendo Aprendizado de Máquina, *Word Embeddings* e *Knowledge Graphs* aplicados à mineração de ideias.

2.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (do inglês *Machine Learning* – ML) pode ser definido como o processo onde determinado algoritmo identifica padrões dos dados e cria previsões sem ser explicitamente programado para fazê-lo (SAMUEL, 2000). É usado para desenvolver algoritmos preditivos complexos e gerar modelos matemáticos mais precisos que analisam grandes quantidades de dados históricos para fazer previsões sobre eventos futuros (JORDAN; MITCHELL, 2015).

A aplicação de técnicas de aprendizado de máquina (ML) em vários campos da ciência vem crescendo rapidamente, especialmente nos últimos 10 anos. Em relação à inteligência artificial (IA), os termos ML e IA podem ser usados como sinônimos. Isso é um equívoco compreensível, já que ML é um subconjunto de IA, mas nem tudo relacionado a IA se enquadra na categoria ML (PADARIAN; MINASNY; MCBRATNEY, 2020).

O ML é um tópico de considerável interesse atualmente na academia e na indústria. Muitas empresas usam algoritmos de aprendizado de máquina para destacar o conhecimento oculto nos dados do consumidor e melhorar os negócios, proporcionar melhor experiência ao cliente e contribuir para uma eficiência operacional como velocidade, economia de custos e maior precisão (BAYOUDE *et al.*, 2018).

Técnicas de aprendizado de máquina oferecem potencial para que sejam obtidos resultados como maior eficácia e eficiência em diferentes tarefas intensivas em conhecimento. Entre essas tarefas destacam-se a classificação (foco nesta tese), a regressão, o agrupamento, a associação, entre outras. Os pontos fortes do aprendizado de máquina incluem a capacidade de lidar com dados de alta dimensionalidade e, no caso da tarefa de classificação, mapear classes com características muito complexas. No entanto, implementar uma classificação com aprendizado de máquina não é simples e a literatura fornece informações conflitantes sobre muitas questões (MAXWELL; WARNER; FANG, 2018).

O aprendizado de máquina é um estudo multidisciplinar de como os computadores usam dados ou experiências anteriores. Com a capacidade de melhorar de forma independente algoritmos específicos, o computador adquire conhecimento por meio da aprendizagem O

aprendizado de máquina é uma das mudanças tecnológicas no mundo moderno da computação que tem um grande impacto em todas as esferas (HOU et al., 2020).

Geralmente, esses métodos tendem a produzir maior precisão em comparação com classificadores paramétricos tradicionais, especialmente para dados complexos, com recursos esparsos e de muitas variáveis, fato este que pode ser comprovado nos trabalhos de Ghimire *et al.* (2012), Loyola-González, Medina-Pérez e Choo (2020), Parcheta *et al.* (2020) e Shuang *et al.* (2020).

A seção a seguir fornece um maior detalhamento sobre a tarefa de classificação de texto utilizando no contexto do aprendizado de máquina.

2.4 CLASSIFICAÇÃO DE TEXTO

A tarefa de classificação de texto surgiu da necessidade natural de organização dos dados em que se objetiva atribuir assuntos semelhantes para determinada classe. Os métodos automáticos, por sua vez, vieram para auxiliar esta tarefa diante do aumento no volume de dados disponíveis para consulta.

Uma das definições de classificação de texto refere-se como o processo de classificar documentos de texto em um número fixo de classes predefinidas (VIJAYAN; BINDU; PARAMESWARAN, 2017). De forma semelhante Altinel e Ganiz (2018), definem a classificação automática de texto como a tarefa de organizar documentos em classes pré-determinadas, geralmente usando algoritmos de aprendizado de máquina.

A classificação de texto define que os objetos são separados em categorias, geralmente para algum propósito específico, onde uma categoria explora uma relação entre as palavras e os seus significados. É uma tecnologia chave para lidar e organizar grandes volumes de documentos, sendo utilizada em aplicações de gerenciamento de informações, alocando automaticamente um documento para uma ou mais classes predefinidas (KADHIM, 2019).

2.4.1 Técnicas de Classificação de Textos

A classificação de texto é usada para extrair conhecimento a partir de padrões do texto não estruturado de várias fontes e tem suas bases na tarefa de classificação da área de Aprendizado de Máquina e técnicas de pré-processamento advindas do Processamento de Linguagem Natural. É uma área de pesquisa que assume o desafio de produzir ferramentas de

inteligência, analisar grandes quantidades de texto em linguagem natural e encontrar padrões (BRINDHA; PRABHA; SUKUMARAN, 2016).

Existem duas formas de realizar a classificação automática de texto: abordagens baseadas em regras e aprendizado de máquina. Na abordagem baseada em regras, as regras de classificação são definidas através de programação e os documentos são classificados com base nessas regras. Esta abordagem promove bons resultados quando o número de regras é pequeno, caso contrário, a manutenção da base de regras se torna difícil à medida que o número de regras aumenta e acabam conflitando entre si (SEBASTIANI, 2002).

Para superar essas limitações, a abordagem de aprendizado de máquina é usada na classificação de texto. De modo geral, objetiva classificar documentos de texto observando as características de um determinado *corpus* e, a partir dessas características, decidir para qual categoria um novo documento desconhecido será atribuído (DWIVEDI; ARYA, 2016).

2.4.1.1 Árvore de Decisão

Uma Arvore de Decisão (do inglês *Decision Tree* - DT) é essencialmente uma decomposição hierárquica do espaço de dados, em que um predicado ou uma condição no valor do atributo é usado para dividir o espaço de dados hierarquicamente. A divisão do espaço de dados é realizada recursivamente na árvore de decisão, até que os nós da folha contenham certo número mínimo de registros (QUINLAN, 1986).

O documento é então classificado para a classe representada pelo nó da folha. Os predicados de decisão nos nós internos podem ser a presença ou ausência dos termos em documentos de texto (VIJAYAN; BINDU; PARAMESWARAN, 2017).

A construção de uma árvore de decisão consiste em partições sucessivas do conjunto de treinamento original em subconjuntos menores. Em um contexto de mineração de texto, mesmo que não sejam necessariamente utilizadas na classificação de novas instâncias, as DTs podem ser construídas para fornecer descrições das características comuns aos membros de cada classe (FRIZZARINI; LAURETTO, 2013).

Os classificadores de árvore de decisão mostram um grande potencial em muitos problemas de reconhecimento de padrões, onde pode-se citar: classificação de dados de várias origens, diagnóstico médico, reconhecimento de fala, entre outros. Uma das principais características das DTs reside na flexibilidade de serem utilizadas com diferentes subconjuntos de recursos e regras de decisão, em diferentes estágios de classificação, bem

como a capacidade de compensações entre precisão de classificação e eficiência de tempo (SAFAVIAN; LANDGREBE, 1991).

Trabalhos como o elaborado pelos autores Ranganathan, Irudayaraj e Tzacheva (2018) utilizam árvore de decisão na área de mineração de emoções e *marketing*. A aplicação desenvolvida inclui dispositivos que detectam as emoções e sugerem recomendações de acordo com os comentários do usuário, aprimorado, segundo os autores, no *marketing* de produtos com impacto no aumento das vendas.

Em Phu *et al.* (2017) é utilizado uma árvore de decisão para classificação semântica de documentos, através da abordagem de classificação de sentimentos. A abordagem demonstrou precisão, porém os autores destacam que o conjunto de treinamento deve ser incrementado para melhorar os resultados.

2.4.1.2 Naive Bayes

Os classificadores *Naive Bayes* (NB) adotam a suposição de que o valor de uma determinada característica é independente do valor de qualquer outra característica em um texto. Na classificação de texto, a suposição em um classificador *NB* refere-se à probabilidade de cada palavra de um documento ser independente da ocorrência de outra palavra no mesmo documento (DENG *et al.*, 2019).

O NB é um algoritmo clássico e tem sido amplamente utilizado na categorização de textos (JOACHIMS, 1997; MCCALLUM; NIGAM, 1998). De modo geral, simplifica muito o aprendizado e compete bem com classificadores mais sofisticados, sendo indicado para situações onde é necessário escolher entre duas condições distintas de uma classe de parâmetros, ou seja, escolhas binárias (RISH, 2001). Seu desempenho competitivo na classificação demonstra bons resultados, uma vez que a suposição inicial de independência condicional em que se baseia, normalmente é verdadeira em uma série de aplicações do mundo real (ZHANG, 2004).

Em Zhang e L1 (2007) os autores utilizaram NB para realizar a detecção de *spam* e mencionam a necessidade de um número elevado de instâncias de treinamento para uma classificação precisa. O mesmo estudo sugere um ajuste dinâmico das probabilidades de ocorrência das palavras (características) durante a classificação para obter um modelo capaz de realizar previsões adequadas.

Embora o algoritmo NB seja caracterizado pela simplicidade, trabalhos recentes como Ababneh (2019) e Chen *et al.* (2019) demonstraram uma boa eficiência na tarefa de classificação de texto em mais de um idioma e com conjuntos de treinamentos pequenos.

2.4.1.3 Máquinas de Vetores de Suporte

Uma Máquina de Vetores de Suporte (do inglês *Support Vector Machine* - SVM) tem como princípio determinar um limiar de separação em um espaço de busca que pode separar adequadamente diferentes classes (AGGARWAL; ZHAI, 2012). Matematicamente, uma SVM gera uma decisão limite que melhor separa duas ou mais classes com uma margem em torno deste limite, que tem sua posição e largura controlada e pode ser vista como um parâmetro utilizado para ajustar a sensibilidade do classificador (CHRISTENSEN *et al.*, 2017b).

Com sua boa capacidade de generalização, as SVMs eliminam a necessidade de seleção de características, tornando a aplicação da categorização do texto consideravelmente mais fácil. Outra vantagem das SVMs sobre os métodos convencionais é a sua robustez. Além disso, as SVMs não exigem nenhum parâmetro de ajuste, pois elas podem encontrar boas configurações de parâmetros automaticamente (JOACHIMS, 1998).

Estudos como o de Coussement e Van den Poel (2008), mostram que as SVMs apresentam bom desempenho de generalização quando aplicadas à mineração de dados utilizando grandes bases de dados e com muitos ruídos. Já Christensen *et al.* (2017b) usaram uma SVM para testar se um classificador de aprendizado de máquina desta natureza poderia aprender o padrão de ideias escritas como texto. A comparação entre o desempenho no conjunto de validação e o desempenho no conjunto de teste demonstra a confiabilidade do classificador utilizado.

2.4.1.4 Florestas Randômicas

Um classificador designado como Floresta Randômica (do inglês Randon Forest - RF) é um método introduzido por Ho (1995) e aprimorado por Breiman (1999) baseado também na classificação de texto, que utiliza uma estrutura composta em árvores.

Os classificadores baseados em árvore podem ter sua capacidade expandida para aumentar a precisão. A essência do método RF é construir várias árvores em subespaços

selecionados aleatoriamente do espaço inicial. Árvores em diferentes subespaços generalizam sua classificação de maneiras complementares, e sua classificação combinada pode ser aprimorada (BREIMAN, 1999).

Porém há um limite que deve ser balanceado para conquistar uma estrutura, pois a medida que mais árvores na RF aumentam, aumenta também a complexidade do tempo de processamento na previsão (KOWSARI *et al.*, 2019).

Um modelo de RF compreende um conjunto de árvores de decisão, cada uma das quais é treinada usando subconjuntos aleatórios de dados. Considerada uma instância, a previsão por RF é obtida pela maioria de votos das previsões de todas as árvores na floresta. Por esta característica, as RFs são adequadas para lidar com os dados ruidosos de alta dimensão na classificação de texto (ISLAM et al., 2019).

No trabalho de Cunningham et al. (2019) é proposto o uso de classificação de texto para detectar *malwares* usando RFs como método de classificação, atingindo bons resultados. Os autores consideram a tarefa de detectar *malwares* como um problema de classificação de texto binário, onde o algoritmo RF apresentou boa precisão nesta tarefa.

O método RF também produziu bons resultados na análise de sentimentos, mais precisamente na classificação de *tweets*. O sarcasmo é a principal razão por trás dos problemas na classificação de *tweets*, visto que esta característica desvia o significado da sua composição real, confundindo a classificação e produzindo resultados falsos. Problema este que foi abordado utilizando RF como classificador tendo apresentado bons resultados (KUMAR; KAUR, 2020).

2.5 WORD EMBEDDING

Word Embedding (*WE*) ou incorporação de palavras é uma forma de representação que permite que palavras com significados semelhantes também tenham uma representação semelhante. Isso possibilita que sistemas computacionais desenvolvam uma melhor compreensão das palavras (AUBAID; MISHRA, 2018). É uma abordagem destinada a capturar as informações semânticas latentes da linguagem, podendo ser utilizada para classificação de texto (MIKOLOV, 2013).

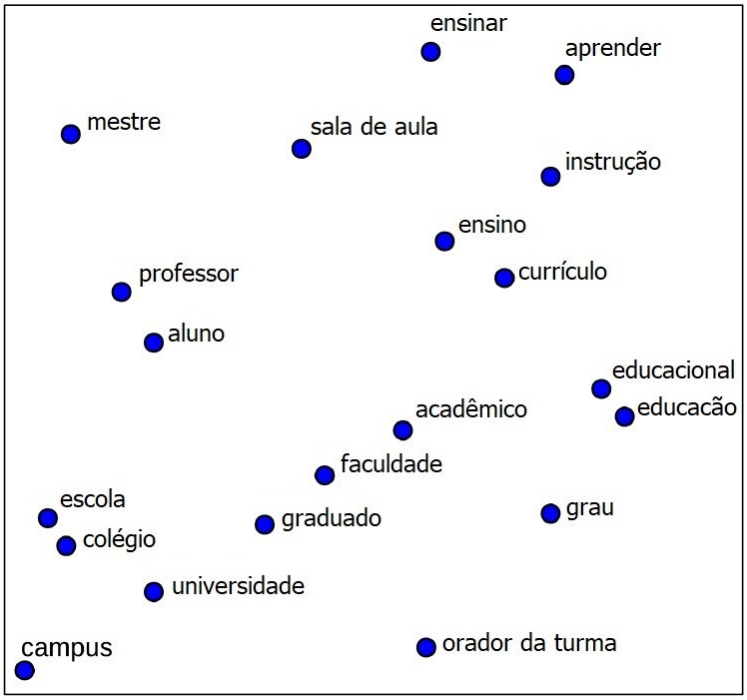
Uma das características do WE é a utilização de vetores densos e de baixa dimensão cujo benefício é o aumento da capacidade computacional, uma vez que não há necessidade de se manipular vetores dispersos e de alta dimensionalidade, como é o caso de alguns outros

classificadores (GOLDBERG, 2013). As representações podem ser facilmente construídas a partir de vetores de incorporação de palavras que também tem a vantagem de não necessitarem de uma grande quantidade de documentos para treinar os modelos (SINOARA *et al.*, 2019).

Os WEs são modelos matemáticos que codificam relações de palavras dentro de um espaço vetorial. Eles são criados por um processo de treinamento baseado em informações de coocorrência entre palavras em uma grande base de informação. É uma ferramenta emergente para o processamento de linguagem natural com aplicação para uma ampla variedade de tarefas de processamento de texto, em que a sua utilidade decorre da capacidade de codificar relacionamentos de palavras no espaço vetorial. As aplicações variam de componentes em sistemas de processamento de linguagem natural até ferramentas para análise linguística no estudo de linguagem e literatura (HEIMERL; GLEICHER, 2018).

A representação WE possui a capacidade de construir uma representação vetorial de palavras, com a propriedade de que palavras semanticamente relacionadas sejam projetadas na mesma vizinhança do espaço (BUTNARU; IONESCU, 2018). Na Figura 1 tem-se um conjunto de palavras representado em um espaço 2D, gerado pela aplicação de um WE de 300 dimensões.

Figura 1 - Exemplo de representação WE de palavras no espaço 2D



Fonte: Traduzido de Butnaru e Ionescu (2018)

Os dois métodos mais conhecidos para implementar WE são *Word2Vec* e *Global Vectors (GloVe)*. Esses dois métodos têm atraído grande atenção e têm sido relatados como os mais eficientes para aprender representações vetoriais de palavras. Por esta razão vêm sendo usados em diferentes tarefas de processamento de linguagem natural (NAILI; CHAIBI; BEN GHEZALA, 2017).

Word2Vec é um método de incorporação de palavras proposto por Mikolov *et al.* (2013). O princípio deste método é aprender vetores dimensionais usando um dos dois modelos neurais distintos: *Continuous Bag of Words (CBOW)* e *Skip-Gram*. O *CBOW* prevê uma palavra atual com base em seu contexto, que corresponde às palavras vizinhas. Já o *Skip-Gram* busca a predição do contexto dado uma palavra.

De acordo com Mikolov *et al.* (2013), cada um desses modelos tem sua própria vantagem. O *Skip-Gram* é mais eficiente com pequenos conjuntos de dados de treinamento e com palavras pouco frequentes. Por outro lado, o *CBOW* é mais rápido e funciona bem com palavras frequentes.

GloVe, outro método conhecido para implementar WE, foi proposto por Pennington, Socher e Manning (2014). É baseado em ocorrências de palavras em um texto, sendo composto de duas etapas principais. A primeira é a construção de uma matriz de coocorrência a partir de um conjunto de treinamento utilizando a frequência da palavra que coocorre com outra palavra. O segundo passo é a fatoração da matriz de coocorrência para obter vetores.

A partir do conceito de WE vários métodos, modelos e aplicações vêm sendo propostos e implementados.

Um modelo de linguagem neural, baseado em tópicos *Skip-Gram*, *Word Embedding* e em Redes Neurais Convolucionais foi proposto por Xu *et al.* (2016) para classificar dados textuais biomédicos, capturando as relações semânticas das palavras com modelos de tópicos. As palavras indexadas no *Word Embedding* são utilizadas como entradas às arquiteturas de Redes Convolucionais Multimodais. Os experimentos conduzidos em vários conjuntos de dados do mundo real mostram que a combinação proposta executa com sucesso as tarefas de classificação de textos, incluindo a indexação de artigos médicos.

Butnaru e Ionescu (2017) propuseram uma nova abordagem para a classificação de texto com base em *Word Embedding*, inspirados em "*Bag of Visual Words*" que é um modelo de palavras, amplamente utilizado em visão computacional. Os autores relatam a obtenção de bons resultados em duas tarefas de mineração de texto, a saber: categorização de texto por tópico e classificação de polaridade.

Um documento pode ser originalmente representado por uma matriz termo-documento, que relaciona todas as palavras e o número de vezes que cada palavra aparece em um documento. No entanto, a matriz termo-documento possui alta dimensionalidade e sua transformação se faz necessário. No trabalho de Hoai Nam e Quoc (2017) os autores analisam duas características de transformação na matriz termo-documento: o modelo baseado na aproximação de baixa classificação (*Low-Rank Approximation*) e o modelo com base em *Word Embedding*. Os resultados da experiência confirmam que o equilíbrio na utilização destes dois recursos cria uma boa condição para o aprimoramento da classificação.

Na sua utilização considerando a semântica das palavras e incorporando a representação implícita do texto, o WE implementado no *Word2vec* foi utilizado na tarefa de classificação de texto baseada no *Open Directory Project*. Ao contrário do uso comum do *Word2Vec*, foram utilizados os vetores de entrada e saída e isso permitiu calcular uma combinação típica de similaridade entre palavras, o que é mais eficaz na classificação do texto (ALIYEVA *et al.*, 2018).

A incorporação de palavras também atua em conjunto com a classificação de texto baseada em regras e o trabalho de Aubaid e Mishra (2018) concentrou-se principalmente nas áreas sociais: ciências, classificação de produtos para compras, bibliotecas digitais e filtragem de *spam*. Os resultados contribuíram para determinar a boa atuação dos sistemas baseados em regras, bem como fornecer um guia para auxiliar os pesquisadores em planejamento de pesquisas futuras.

Em Kilimci e Akyokus (2018) foram utilizadas diferentes representações de documentos com o WE e um conjunto de classificadores de base para classificação de texto. O conjunto de classificadores de base inclui algoritmos de aprendizado de máquina, como *Naive Bayes*, *Support Vector Machine*, *Random Forest* e uma *Deep Learning-Based Convolutional Network*. Foi realizada a análise da precisão da classificação de diferentes representações de documentos empregando um conjunto de classificadores e os resultados experimentais mostraram que o uso de métodos de aprendizado profundo em conjunto com WE melhora o desempenho da classificação dos textos. O autor cita como classificadores de melhor desempenho o *Random Forest* e a *Deep Learning-Based Convolutional Network*.

Um método geral de pré-processamento então foi proposto para cenários em que os dados de treinamento são escassos. Ele agrupa termos semanticamente semelhantes através do WE, que simulam como humanos pré-processam textos, substituindo palavras desconhecidas

por termos conhecidos e também agrupando palavras semanticamente semelhantes (ELEKES *et al.*, 2019).

Em Guo *et al.* (2019) é proposto um novo esquema de ponderação de termos combinados com WE para melhorar o desempenho da classificação das redes neurais convolucionais. Primeiramente, é executado o método de WE para mapear cada palavra em um vetor de palavras e depois foi utilizada uma rede neural convolucional para realizar a classificação.

Uma estratégia de agrupamento hierárquico modificada por WE pré-treinada também foi proposta para classificação de texto, em um modo de aprendizado por transferência de domínios. O modelo aproveita e transfere o conhecimento obtido de alguns domínios de origem, para reconhecer e classificar as sequências de texto a partir de exemplos no domínio do problema de destino (PAN *et al.*, 2019).

Um novo método de representação de texto denominado *Hybrid Word Embeddings* foi proposto por Song, Srimani e Wang (2019) combinando informações semânticas e contextuais obtidas do *WordNet®* *e* extraídas de documentos, para fornecer representações concisas e precisas dos textos. O estudo experimental sobre classificação de documentos mostra que o método proposto supera os métodos existentes, incluindo o *Word2Vec*, em termos de precisão de classificação.

O trabalho de Stein, Jaques e Valiati (2019) é um estudo investigativo sobre a aplicação de modelos e algoritmos sobre o problema específico da classificação hierárquica de texto. Foram treinados modelos de classificação com destaque em implementações de algoritmos de aprendizado de máquina: *fastText*, *XGBoost*, *Support Vector Machines* e redes convolucionais utilizando a biblioteca *Keras®*, bem como, métodos de incorporação de palavras como o *GloVe* e o *Word2Vec.* Os resultados foram avaliados com medidas apropriadas para o contexto hierárquico. O *fastText* alcançou o melhor desempenho em conjunto com a incorporação de palavras e demonstrou ter uma abordagem muito promissora para a classificação hierárquica de texto.

Em Topaz *et al.* (2019) foi aplicado o sistema *NimbleMiner* em uma grande amostra de anotações clínicas de pacientes internados para identificar casos de diabetes, apresentando bons resultados com o uso de incorporação de palavras no domínio da saúde. O *NimbleMiner* é um sistema de processamento de linguagem natural baseado em WE, para classificação de textos clínicos, independentemente do idioma.

2.6 KNOWLEDGE GRAPH

Um *Knowledge Graph* (KG) ou grafo de conhecimento, representa a espinha dorsal de muitos sistemas de informação que requerem acesso à estruturas de conhecimento, seja de domínio específico ou domínio independente (PAULHEIM, 2016). KG representa o conhecimento humano no formato legível por máquinas e estão se tornando a base importante de muitas aplicações em áreas como a inteligência artificial e o processamento de linguagem natural (WANG *et al.*, 2018)

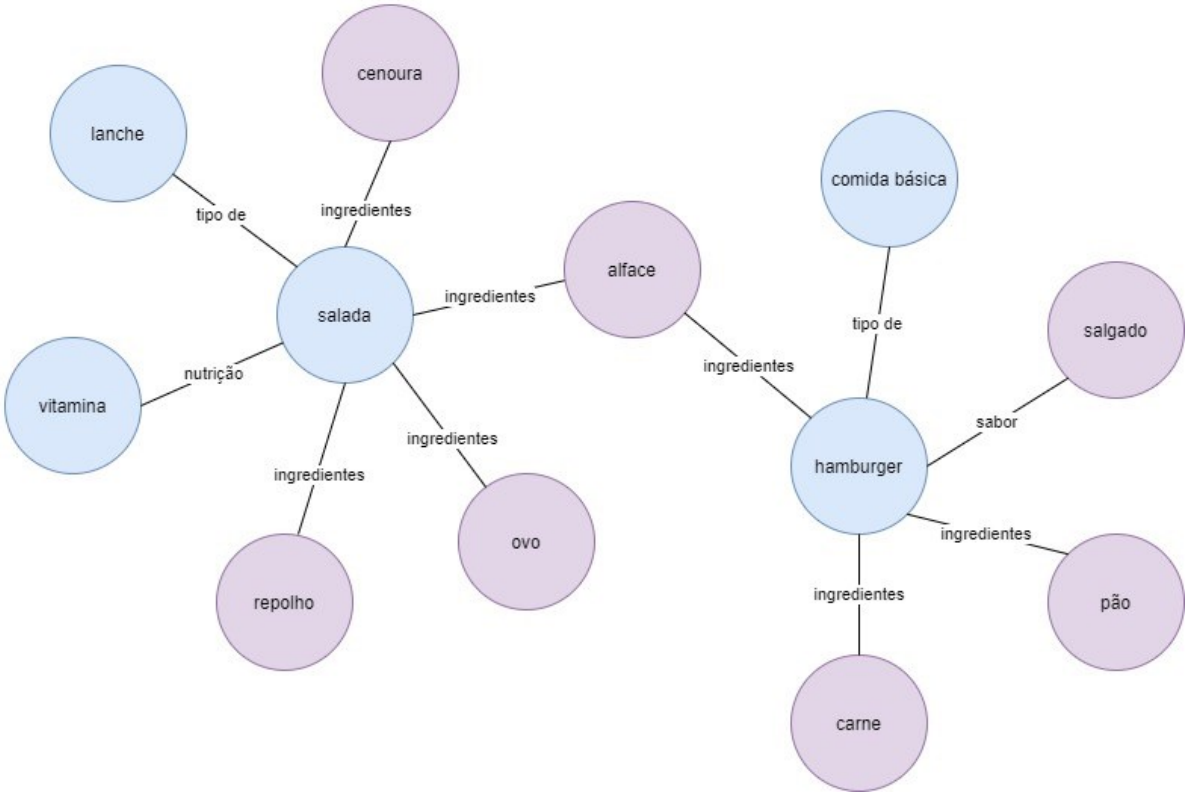
Um KG é um grafo direcionado, com entidades do mundo real como nós e suas relações como arestas. Neste grafo, cada aresta direcionada, juntamente com duas entidades, a principal e a cauda, constituem uma tripla, ou seja, entidade principal, predicado e entidade cauda, que também é chamado de fato. O conhecimento do mundo real quando modelado pelos KGs costumam conter milhões ou bilhões de fatos (LIN *et al.*, 2015; WANG *et al.*, 2017).

Devido aos KGs terem a função de ligar a variável alvo ou atributos da variável com o conteúdo dos dados de texto (por exemplo, entidades e relações mencionadas no texto), podem ser utilizados para construir recursos eficazes voltados, por exemplo, às tarefas de previsão e classificação (JIANG; ZHAI; MEI, 2018).

Recentes esforços nas áreas de extração de informações e engenharia de conhecimento criaram KGs em grande escala, em que múltiplas entidades e relações são extraídas a partir de dados não estruturados e semiestruturados, sendo posteriormente processados manualmente ou automaticamente e, em seguida, organizados em um KG (BOLLACKER *et al.*, 2008; ROTMENSCH *et al.*, 2017).

Os grafos de conhecimento podem efetivamente organizar e representar o conhecimento para que ele possa ser eficientemente utilizado em aplicativos e sistemas. O raciocínio de conhecimento sobre gráficos de conhecimento visa inferir novas conclusões a partir dos dados existentes. A partir da análise do KG e de suas relações, informações podem ser derivadas por meio de raciocínio de conhecimento e pode fornecer dados para apoiar os aplicações (CHEN; JIA; XIANG, 2020). Na Figura 2 é mostrado um KG obtido através de dados do domínio alimentar onde é mostrada a relação construída entre os elementos.

Figura 2 - Exemplo de Grafo de Conhecimento



Fonte: Traduzido de Yu *et al.* (2020)

Existem diversos trabalhos publicados onde os KGs são utilizados como classificadores, dentre os quais serão citados alguns, nos parágrafos a seguir.

Em Marin *et al.* (2014) é proposto um método para aprender padrões de frases em um classificador de texto, usando informações externas na forma de um KG. O estudo mostrou que a combinação entre a utilização de classificadores e *knowledge graph* obteve um bom desempenho e apontou que melhorias no conjunto de treinamento podem tornar o sistema mais assertivo.

No artigo de Wang, Guo e Liu (2019) é proposto um método para aumentar a eficácia de uma base de dados treinada com incorporação de palavras por meio do KG, para fornecer uma representação mais adequada de vetores de palavras faltantes voltada à classificação de textos. Esta técnica tem como objetivo preencher o número de vetor de palavras faltantes no mapeamento do vetor de palavras de um classificador de textos, fato este que é comum em alguns idiomas.

Experimentos comprovaram que a introdução de conhecimento prévio do KG pode melhorar o desempenho da classificação de texto. Combinando KGs com redes neurais para

melhorar a informação granular promove uma aprendizagem profunda da rede neural, proporcionando tal melhora. Nos modelos de aprendizagem é importante utilizar com eficácia uma grande quantidade do conhecimento existente e reduzir a dependência em amostras para melhorar a eficiência da classificação de texto (MENG; WANG; LIU, 2019).

O artigo de Jiang *et al*. (2020) propõe um novo método baseado em KG utilizado para classificação de textos curtos, chamado *BaKGraSTeC*. Este método emprega conhecimento externo ao KG para enriquecer as informações de texto e utilizar suas informações estruturais por meio de uma rede neural, para promover a compreensão do texto. Além disso, é inserido um mecanismo que emite avisos considerando semelhança e coocorrência entre conceitos encontrados em textos.

Em Tosi e Reis (2021) é proposto o *SciKGraph*, um *framework* para estruturar o conhecimento de um campo científico considerando a semântica dos conceitos extraídos de documentos textuais da área, para construir grafos de conhecimento. A abordagem agrupa os grafos de conhecimento em seus principais tópicos e extrai automaticamente informações, como os conceitos mais relevantes e sobreposição de conceitos entre os tópicos. De modo geral, *SciKGraph* contribui estruturando o conhecimento que pode auxiliar pesquisadores no estudo de suas áreas, reduzindo o esforço e o tempo dedicado à tarefas que envolvem a sintetização de informações.

2.7 TRABALHOS CORRELATOS NA ÁREA DE MINERAÇÃO DE IDEIAS

O termo mineração de ideias só foi introduzido na literatura no ano de 2010 por Thorleuchter; Van Den Poel e Prinzie (2010) como “um processo de extrair ideias novas e úteis a partir de texto não estruturado usando métodos de mineração de texto”. Porém, em trabalhos anteriores a 2010, já era introduzido o mesmo conceito, porém utilizando outros termos.

Outro trabalho que introduz o conceito de mineração de ideias, porém no contexto organizacional, mas sem utilizar explicitamente o termo, é o de Paukkeri (2009). Este estudo propõe a recuperação de ideias dentro do local de trabalho, onde normalmente os colaboradores tendem a escrever suas ideias e invenções em cadernos ou armazená-los em arquivos de seus *laptops* e telefones celulares. Posteriormente, com certa frequência torna-se difícil encontrar a anotação necessária ou criar uma visão geral de todas as ideias, isto sem mencionar os numerosos pedaços de papel perdidos. Para realmente apoiar a criatividade e

inovações nas organizações, devem existir ferramentas de *software* capazes de auxiliar na coleta e gerenciamento de ideias. Ferramentas estas que devem fornecer uma maneira simples de inserir notas e facilitar o posterior acesso às ideias produzidas por qualquer integrante da organização.

Thorleuchter *et al.* (2010) introduz o termo mineração de ideias e descreve este processo em três etapas. A primeira etapa se concentra no problema, o usuário tem que fornecer informações textuais onde ele descreve um problema específico. Na segunda etapa o usuário deve fornecer mais informações textuais que, provavelmente, possam resolver o seu problema. Finalmente, a terceira etapa consiste em avaliar todos os padrões de texto extraídos quanto à novidade e utilidade. Isso significa que eles são comparados à descrição do problema usando uma medida de mineração de ideia específica capaz de classificar os padrões retornados como uma ideia nova e útil.

Em 2011 surgem trabalhos utilizando mineração de ideias aplicada a domínios específicos. A definição da ideia não se restringe apenas ao domínio tecnológico, pode também ser utilizada no domínio do comportamento social. Porém, essas ideias de diferentes domínios precisam de parâmetros diferentes para que a extração e a identificação de novas ideias possam ser consideradas bem sucedidas (THORLEUCHTER; HERBERZ; POEL, 2011)

A aplicação da mineração de ideias no domínio tecnológico teve sua comprovação de sucesso a partir de um estudo de caso apresentado por Thorleuchter e Van Den Poel (2012). Este trabalho também menciona a grande quantidade de informação textual acessível na *internet*, abrangendo diferentes tópicos. Esta informação pode ser uma fonte valiosa para os tomadores de decisão, pois podem conter ideias interessantes, que possivelmente são relevantes para resolver problemas que envolvam a tomada de decisão.

Em 2012 tem-se o registro do trabalho de Tripathy *et al.* (2012) que combina mineração de ideias com buscas na *internet*. O processo de busca de ideias na *internet* se inicia com um usuário fornecendo uma descrição do produto. Em seguida, a descrição do produto fornecida é separada usando uma lista padrão de palavras (termos) relevantes. As ideias identificadas devem ter duas propriedades: novidade e utilidade. Novidade significa que a descrição da ideia contém informações que não devem estar na descrição do produto. Utilidade significa que a descrição da ideia também contém informação que deve estar na descrição do produto. E para haver uma ideia nova e útil, deve haver uma combinação entre estas duas propriedades.

Thorleuchter e Van Den Poel (2013) propõem uma nova metodologia que permite uma identificação de ideias a partir da *internet* capaz de resolver determinado problema. A etapa de mineração da *web* é usada para identificar *sites* da *internet* onde ideias relevantes são descritas. Para acessar essas informações textuais, são usados os mecanismos de busca tradicionais.

O processo da abordagem de mineração na *web* é baseado no mesmo processo da abordagem de mineração de ideias onde o usuário fornece um contexto (descrição do problema) e, após o pré-processamento, são criados os vetores de termos e as consultas de pesquisa representando o problema. As consultas são executadas por meio de um mecanismo de pesquisa da *web* e os vetores de termos são criados a partir dos resultados da consulta com base nas características da ideia para um domínio em particular. O tamanho do vetor é definido pelo número de palavras-chave obtidas no texto, onde se supõe a existência de uma nova ideia. Esses vetores são comparados a vetores fornecidos na descrição do problema, utilizando-se uma medida euclidiana que calcula a distância, e com isso se obtém uma proposta de medida de mineração de ideias. Como resultado, ideias úteis são extraídas da *internet* (THORLEUCHTER; VAN DEN POEL, 2013)

Uma consideração relevante foi apresentada por Li e Chen (2014) que, de certa forma, faz um alerta aos trabalhos publicados. Na mineração de ideias os pesquisadores se concentram em como explorar algoritmos para extrair, a partir de textos, padrões implícitos desconhecidos e potencialmente úteis, mas ignoram os componentes do conhecimento desses padrões. O conhecimento ou padrões ocultos descobertos pela mineração de ideias podem ser chamados de “conhecimento aproximado”. Esse conhecimento deve ser posicionado em um domínio para então derivar o conhecimento aceito pelos usuários ou organizações, para então ser chamado de “conhecimento inteligente”.

Seguindo a mesma linha de trabalhos que mencionam a avaliação das ideias coletadas por métodos computacionais, Klein e Garcia (2015) afirmam que as técnicas de mineração de ideias são fundamentalmente limitadas. Isto pelo fato de que algoritmos atuais de processamento de linguagem natural têm apenas uma compreensão superficial da linguagem natural e, portanto, podem ser facilmente enganados. Neste sentido, os autores propõem a utilização de avaliadores para a análise de ideias candidatas com base em uma descrição clara dos critérios de seleção estabelecidos, e assim, realizarem uma filtragem adicional.

Por outro lado, os métodos de mineração de ideias vêm evoluindo e se tornando mais eficazes. Liu *et al.* (2015) propõem um método de extração de ideias a partir da análise de

resumos de artigos, mais especificamente explorando os títulos e resumos. O método consiste em descobrir as ocorrências de soluções de problemas entre palavras e frases detectadas nos resumos e classificar as ideias obtidas como potencialmente inovadoras.

Ainda no contexto de evolução e aprimoramento, a mineração de ideias começa a incorporar conceitos já conhecidos nas áreas de administração e planejamento estratégico, os quais propõem a classificação de eventos em sinais fortes (eventos conhecidos) e sinais fracos (eventos incertos e inesperados). A abordagem proposta por Thorleuchter e Van den Poel (2015) tem como objetivo identificar sinais fortes e fracos emergentes que possuem relevância para um determinado problema de decisão estratégica da organização. Esta etapa de identificação de sinais é utilizada para filtrar os resultados obtidos na mineração de ideias.

Em Alksher *et al.* (2016) foi apresentada uma revisão de métodos de mineração de ideias. Este trabalho apresenta de forma resumida a explicação de várias técnicas e conceitos de mineração de ideias já utilizadas. Como principal contribuição destaca-se a constatação de que a mineração de ideias é uma área de pesquisa aberta, que ainda apresenta lacunas em caracterizar a definição de ideia, por se tratar de um conceito bastante amplo e com várias interpretações.

Muitas vezes as ideias, que são usadas como ponto de partida para uma pesquisa são de natureza interdisciplinar, pois combinam aspectos de diferentes disciplinas. A identificação de ideias interdisciplinares numa fase precoce possibilita a introdução de pesquisas mais abrangentes que permitem avanços no processo de inovação. A possibilidade de que uma inovação seja bem-sucedida é normalmente maior aplicando-se este tipo de pesquisa. Porém, uma ideia que hoje é considerada interdisciplinar possivelmente pode ser considerada como disciplinar futuramente. Assim, a natureza interdisciplinar de uma ideia depende das disciplinas e de suas definições atuais (THORLEUCHTER; VAN DEN POEL, 2016).

Uma abordagem ligeiramente diferente dos trabalhos anteriores consta em alguns estudos a partir de 2016, que consiste na utilização de algoritmos de aprendizagem combinados com a mineração de ideias. O trabalho de Christensen *et al.* (2017a) propõe uma seleção aleatória de mensagens (três mil) por dois avaliadores de ideias. Os avaliadores foram instruídos a lerem cada texto e avaliarem se estes textos continham sugestões sobre produtos, melhorias, ou oportunidades de negócios, classificando-os com pesos. A partir disso, os dados representaram a entrada para algoritmos de aprendizado de máquina. Os resultados mostraram que é possível a utilização de inteligência artificial e aprendizagem de máquina para aprender e reconhecer conceitos abstratos como ideias.

Ainda em continuidade com o trabalho acima, Christensen *et al.* (2017b) propuseram um sistema que utiliza como entrada uma grande quantidade de textos de ideias relevantes e não relevantes e, dessa forma, o sistema depois de treinado identifica essas diferenças. A principal conclusão reside na possibilidade de um classificador genérico baseado em aprendizado de máquina ser usado para detectar ideias em qualquer contexto.

O trabalho de Lee e Tan (2017a) explora e rastreia a dinâmica do desenvolvimento de ideias que consiste em três fases: descoberta, identificação e análise de ideias. Ainda na mesma linha de atuação, Lee e Tan (2017b) continuaram o processo de análise de ideias com foco na construção de conhecimento, embora tenham classificado o processo como desafiador devido à amplitude e complexidade dos dados. Este trabalho atuou na análise de aprendizagem, propondo um método que combina análise temporal e aprendizado de máquina não supervisionado para analisar ideias e investigar a natureza das mesmas.

O modelo apresentado no trabalho de Sérgio *et al.* (2017) objetiva auxiliar no processo de tomada de decisão no domínio de Gestão de Ideias. Para atingir o objetivo foi desenvolvida uma ontologia de domínio voltada à representação semântica das ideias e, a partir disso, aplicou-se a análise de agrupamento. Através da análise de agrupamentos foi possível evidenciar padrões e tendências com relação às ideias analisadas, classificando-as como potenciais ideias.

O trabalho de Alksher *et al.* (2017) está voltado para a avaliação dos resultados da mineração de ideias. Este propõe um método para comparar diferentes escalas de medida, bem como investigar a validade e confiabilidade do julgamento humano como critérios de avaliação, utilizando uma análise estatística. Sua principal contribuição foi demonstrar que o uso da escalas de avaliação de ideias são limitadas e portanto, não devem ser usadas para a avaliação de ideias.

A mineração de ideias também demonstrou sua aplicação na busca por ideias relevantes a partir de artigos científicos, mais especificamente analisando os seus resumos. Os resumos científicos são semelhantes em termos de sua estrutura de informação. Alguns dos resumos podem conter o objetivo do estudo, os métodos utilizados no estudo e a conclusão que descreve os resultados obtidos (GUO *et al.*, 2010). No trabalho de Alksher *et al*. (2018a) assume-se que a ideia pode ser identificada dentro do resumo de um artigo. Como tal, o resultado é um método de identificação de ideias que sugere qual parte do resumo contém a principal ideia discutida no artigo reduzindo a complexidade da busca e tornando o método mais assertivo.

Em continuidade ao trabalho mencionado anteriormente, Alksher *et al.* (2018b) propõem melhorias ao seu modelo de mineração de ideias considerando também as relações semânticas entre termos baseados em sinônimos. Os resultados desta investigação mostraram um aumento da eficiência do método em comparação com o modelo apresentado em seu trabalho anterior.

Ainda na mesma linha da utilização de mineração de ideias aplicada a artigos, Azman *et al.* (2019) utilizaram inteligência artificial e métodos de ajustes de curvas para descobrir a posição do texto do resumo de artigo com maior probabilidade de conter a ideia. O resumo é igualmente dividido em três seções: introdução, corpo e conclusão. Verificou-se que a maioria das ideias está localizada na introdução ou na conclusão de um resumo.

Em Sérgio e Gonçalves (2019) é proposto um modelo de mineração de ideias capaz de contribuir na análise e interpretação dos dados coletados em plataformas de Gestão de Ideias, com o intuito de auxiliar no processo de tomada de decisão. A característica importante do modelo proposto é a presença de uma ontologia responsável por representar o conhecimento de domínio.

As comunidades onde usuários compartilham voluntariamente ideias sobre novos produtos ou serviços tem se tornado uma importante fonte de inovação. Em função do volume de informações textuais nessas comunidades, Kim e Park (2019) propõem uma abordagem integrada, construindo duas bases de dados sobre ideias versus produtos existentes usando mineração de texto e verificando a sobreposição entre os dois usando raciocínio baseado em casos. Essa abordagem permitiu identificar oportunidades de inovação e produtos de referência para adaptação. O uso da mineração de texto aumentou a variedade de potenciais oportunidades a serem aproveitadas nas comunidades de inovação. Já a eficácia da identificação foi incrementada através de raciocínio baseado em casos realizando uma verificação de sobreposição entre ideias e produtos ou serviços existentes.

O objetivo do trabalho de Röltgen *et al.* (2020) foi desenvolver e avaliar uma ferramenta organizacional colaborativa e gratuita chamada *IdeaCheck*. Usando um estudo de caso para explicar e relatar experiências com a ferramenta durante a primeira implementação, esta possibilita ao usuário discutir as etapas que uma ideia deve ser avaliada até que seja aceita. Os critérios de avaliação também são registrados, o que torna o processo de escolha transparente aos usuários.

A mineração de ideias foi também aplicada para avaliação de ideias criadas por alunos de cursos *online*. A partir de textos produzidos por esses alunos, o critério utilizado para

selecionar ideias consistiu em buscar os textos que continham a combinação de meios desconhecidos e fins conhecidos, ou em outras palavras, se um fim foi atingido utilizando um meio novo, caracteriza então a presença de uma ideia (AZMAN *et al.*, 2020).

Em Ayele (2020) é proposto um modelo reutilizável de mineração de ideias, o *CRISP-IM*. O *CRISP-IM* pode ser usado para guiar o processo de identificação de tendências usando conjuntos de dados de literatura acadêmica, patentes organizadas temporariamente ou qualquer outro conjunto de dados textuais de qualquer domínio para extrair ideias, através de aprendizado de máquina não supervisionado e análise estatística.

Em Ayele e Juell-Skielse (2020) é apresentada uma melhoria do modelo *CRISP-IM* dividindo-o em camadas. A primeira apresenta a camada de negócios, onde são detalhadas as tarefas executadas por escalas de tecnologia, incubadoras, aceleradoras, consultores e gerentes de concurso. A segunda apresenta a camada técnica onde as tarefas são realizadas por cientistas de dados, engenheiros de dados e especialistas.

Uma outra revisão sistemática da literatura na área de mineração de ideias foi publicada onde se lista as principais técnicas utilizadas para a realização de mineração de ideias. Sendo elas recuperação de informação, inteligência artificial, aprendizado profundo, aprendizado de máquina, técnicas estatísticas, processamento de linguagem natural (PLN) e análise morfológica baseada em PLN (AYELE; JUELL-SKIELSE, 2021).

O estudo de Ozcan *et al.* (2021) teve como objetivo criar um modelo de classificação para identificar *tweets* que continham uma ideia para explorar tendências e recuperar ideias para vários fins. As contribuições práticas se aplicam a inovação, gestão, desenvolvimento de produtos e sustentabilidade.

E, por fim, tem-se o estudo de Ha e Geum (2022), que sugere uma estrutura para identificar novos serviços usando uma abordagem de identificação de palavras-chave específicas para construir uma matriz que expressa a relação entre estas palavras. Foi utilizado o algoritmo de agrupamento *k-means* (um algoritmo de clusterização ou agrupamento) para a identificação de palavras-chave e os autores apontam que os resultados do estudo são promissores e refletem diretamente as necessidades do mercado.

2.8 CONSIDERAÇÕES FINAIS

A fundamentação teórica detalhada nesta seção abrange os conceitos utilizados na proposição e desenvolvimento do modelo desta tese.

Os conceitos de mineração de texto e mineração de ideias estão relacionados com os objetivos geral e específicos da tese e fornecem os conceitos introdutórios para a compreensão do tema. Com esta finalidade são detalhados os conceitos de ideia, bem como os critérios utilizados por especialistas no processo de identificação e seleção de ideias.

O aprendizado de máquina por sua vez representa um ponto central na construção do modelo, pois procura simular o comportamento humano através de algoritmos computacionais.

Na seção sobre classificação de texto são apresentadas as técnicas de classificação de texto que foram utilizadas no modelo, baseadas em uma revisão da literatura onde foram elencadas e detalhadas técnicas amplamente utilizadas nesta tarefa.

São apresentadas ainda as técnicas *Word Embeddings* e *Knowlwdge Graph*, que têm sua utilização em várias áreas e que podem ser também utilizadas como classificadores de texto, porém incorporando ao modelo estruturas de conhecimento.

Finalmente, discute-se um conjunto de trabalhos relacionados à mineração de ideias, através de uma classificação cronológica, onde se pode ter uma noção da evolução sobre o assunto no decorrer dos anos.

3 METODOLOGIA DA PESQUISA

Este capítulo apresentará como a pesquisa foi realizada, classificando-a quanto à natureza, abordagem e procedimentos adotados. Será realizada a descrição quanto à metodologia adotada, no caso o *Design Science Research* (DSR), sendo apresentados os detalhes da revisão sistemática realizada e, por fim, o detalhamento da pesquisa indicando os procedimentos metodológicos adotados caracterizados nesta tese como desenvolvimento da pesquisa.

3.1 ENQUADRAMENTO METODOLÓGICO

Considerando a visão de mundo proposta por Morgan (1980) a pesquisa se enquadra no quadrante inferior direito do modelo de Morgan, caracterizando-se como funcionalista. A perspectiva funcionalista é essencialmente reguladora e prática e está interessada em compreender a sociedade de maneira que produza conhecimento empírico útil. O funcionalismo se faz pelo fato do cientista engajar-se totalmente com a ciência e por meio dela a realidade ser observada por uma lente objetivista (BURRELL; MORGAN, 1979).

Como modalidade de pesquisa, a proposta se enquadra como pesquisa tecnológica, tendo como objetivo a criação de um artefato tecnológico. Segundo Aguiar (1991) a pesquisa tecnológica é o trabalho sistemático, delineado a partir de conhecimento preexistente, obtido através da pesquisa científica e/ou da experiência prática, e aplicado na produção ou aperfeiçoamento de produtos, processos ou serviços.

O objeto da pesquisa tecnológica é o conhecimento prescritivo, uma vez que o artificial "constitui-se em um sistema adaptado ao ambiente em função de determinado propósito humano, um objeto (artefato) com propriedades desejadas, idealizado e fabricado conforme desenho e projeto (*design*)" (CUPANI, 2011).

Do ponto de vista da sua natureza, pode ser considerada uma pesquisa aplicada, pois objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos (SILVA; MENEZES, 2005). A pesquisa aplicada é a investigação original concebida pelo interesse em adquirir novos conhecimentos. É, entretanto, primordialmente dirigida em função de um fim ou objetivo prático específico (OCDE, 2013).

O referencial teórico será obtido através de uma visão interdisciplinar, com objetivo de estudar os métodos e técnicas da Engenharia do Conhecimento e suas relações com a

Mineração de Ideias, auxiliando no processo de identificação de ideia. Também são explorados os critérios de seleção de ideias, de forma a promover subsídios para a construção do modelo. O instrumento empregado constitui-se na pesquisa bibliográfica com foco na revisão sistemática dos constructos que suportam esta tese por meio de bases de dados internacionais.

3.2 DESIGN SCIENCE RESEARCH METHODOLOGY

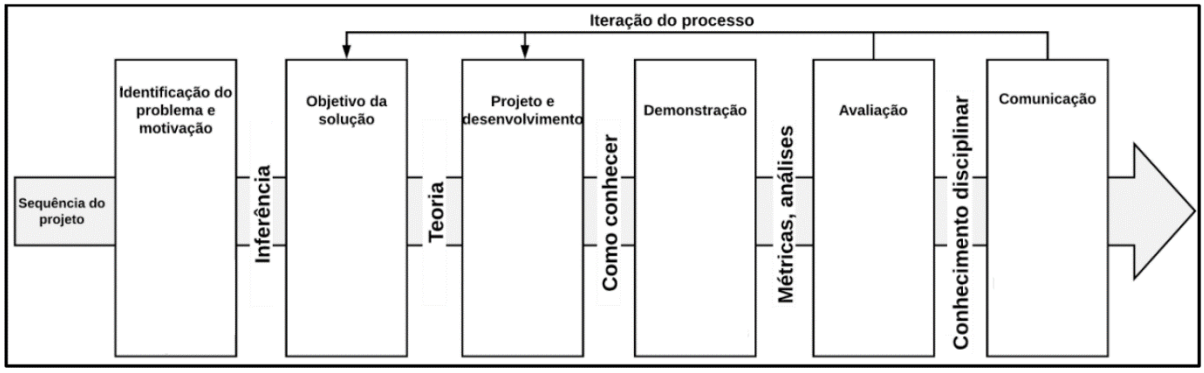
A metodologia *Design Science Research Methodology* (DSRM) serviu de base para esta pesquisa, mais especificamente a proposição de condução de Peffers *et al.* (2007). Segundo Peffers *et al.* (2007) a DSRM é desenvolvida a partir de seis etapas que podem ser executadas sequencialmente ou de acordo com a necessidade de projeto:

- Etapa 1 - Identificação do problema e sua motivação: esta etapa é dedicada à definição do problema de pesquisa específico, apresentando-se uma justificativa para a sua investigação. É importante que a definição deste problema seja empregada na construção de um artefato que pode efetivamente oferecer a solução para este problema. Tem-se como recursos necessários para esta etapa o estado da arte do problema e da relevância da solução apresentada.
- Etapa 2 - Definição dos objetivos para a solução: tendo-se como ponto de partida o conhecimento acerca do problema, bem como a noção do que é viável e factível, delineiam-se os objetivos da solução a ser desenvolvida. São elencados como requisitos desta etapa novamente o estado da arte do problema e o conhecimento das possíveis soluções já previamente apresentadas.
- Etapa 3 - Projetar e desenvolver: etapa destinada à criação do artefato, determinando-se a sua funcionalidade desejada para o artefato, sua arquitetura e em seguida a criação do próprio artefato. Os recursos necessários para a terceira etapa compreendem o conhecimento da teoria que pode ser exercida em uma solução.
- Etapa 4 – Demonstração: momento de demonstração do uso do artefato resolvendo uma ou mais instâncias do problema por meio de um experimento ou simulação, estudo de caso, prova formal ou outra atividade apropriada. Os recursos relacionados para esta etapa incluem o conhecimento efetivo de como usar o artefato para resolver o problema.

- Etapa 5 – Avaliação: nesta etapa deve-se observar e mensurar como o artefato atende à solução do problema, comparando-se os objetivos propostos para a solução com os resultados advindos da utilização do artefato. Nesta etapa pode-se definir pela recursividade da metodologia, isto é, o retorno às etapas 3 ou 4, de modo a aprimorar o artefato.
- Etapa 6 – Comunicação: momento de divulgação do problema e da relevância da propositura de uma solução para o mesmo, além da apresentação do artefato desenvolvido.

A Figura 3 detalha graficamente as etapas da metodologia DSR descritas acima.

Figura 3 – Etapas da DSRM



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

3.3 REVISÃO SISTEMÁTICA DA LITERATURA

A revisão sistemática consiste em um tipo de investigação científica que tem por objetivo reunir, avaliar criticamente e conduzir uma síntese dos resultados de múltiplos estudos primários (COOK; MULROW; HAYNES, 1997). Ela também objetiva responder a uma pergunta claramente formulada, utilizando métodos sistemáticos e explícitos para identificar, selecionar e avaliar as pesquisas relevantes, coletar e analisar dados de estudos incluídos na revisão (CLARKE; HORTON, 2001).

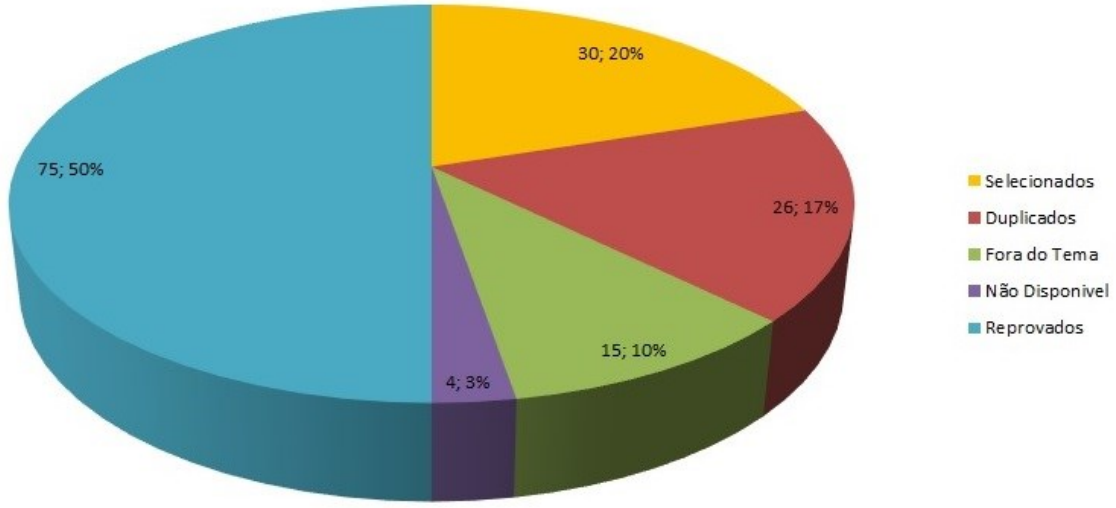
Os instrumentos empregados neste trabalho foram a pesquisa bibliográfica com foco na revisão sistemática do tema "mineração de ideias", utilizando bases de dados internacionais, entre elas: Scopus®, Science Direct®, ACM®, IEEE® e Springer Link®. A escolha destas bases de dados ocorreu pelo fato de serem os principais meios de publicação do tema em pesquisa e, segundo Almeida (2006), interdisciplinares e internacionais.

Os critérios de busca foram utilizados com o termo "*idea mining*" sendo este o conceito central da pesquisa. Para captar artigos que tenham o contexto de aplicação prática, o termo "*idea mining*" foi combinado com os termos "*idea management*", "*data mining*", "*text mining*", "*machine learning*", "*artificial intelligence*" e "*knowledge representation*". Desta forma, a pesquisa foi realizada com a expressão de busca a seguir:

- ("idea mining") AND ("idea management" OR "text mining" OR "data mining" OR "machine learning" OR "artificial intelligence" OR "knowledge representation")

A busca retornou inicialmente 150 artigos distribuídos da seguinte maneira: ACM® (42), Science Direct® (35), Scopus® (38), Springer Link® (21) e IEEE® (14). Após a eliminação das duplicações (26 ao todo), foi realizada a análise inicial com os 124 artigos restantes. Através da leitura dos títulos verificou-se que 15 não possuíam relação direta com os temas abordados nessa pesquisa, sendo então removidos. Após efetuar a leitura dos resumos também foram excluídos 75 artigos, também por não possuírem relação direta com os temas abordados. Foram ainda descartados 4 artigos que somente estavam disponíveis mediante pagamento. Após a análise restaram 30 artigos para serem analisados. A Figura 4 apresenta os percentuais dos artigos obtidos na revisão sistemática.

Figura 4 – Porcentagens dos artigos obtidos na revisão sistemática



Fonte: autor

Não houve restrição temporal na busca realizada, recuperando publicações do período compreendido entre o ano de 2009 e o ano de 2022 A Figura 5 apresenta a série histórica dos artigos recuperados e selecionados para esta revisão sistemática.

Figura 5 - Série histórica dos artigos recuperados e selecionados na revisão sistemática

Fonte: autor

A partir do conjunto de artigos selecionados para a revisão sistemática, foi possível identificar uma lista de 143 palavras-chave e a partir da lista foi produzida uma nuvem de *tags* apresentada na Figura 6.

Figura 6 - Nuvem de *tags* produzida a partir das palavras-chave dos artigos

Fonte: autor

Nos 30 estudos analisados, observou-se uma lista de autores e instituições: são 54 autores, filiados a 29 instituições de 17 países. O Quadro 1 apresenta os autores por ordem de quantidade de publicações, com os seus respectivos números de artigos publicados sobre o tema, afiliação e país.

Quadro 1 - Autores por ordem de quantidade de publicações

Autor	**Número de Estudos**	**Afiliação**	**País**
Dirk Thorleuchter	6	Fraunhofer INT	Alemanha
Dirk Van den Poel	6	Ghent University	Bélgica
Mostafa A. Alksher	6	Universiti Putra Malaysia	Malásia
Azreen Azman	6	Universiti Putra Malaysia	Malásia
Razali Yaakob	6	Universiti Putra Malaysia	Malásia
Eissa M. Alshari	6	IBB University	Iémen
Rabiah Abdul Kadir	4	Universiti Kebangsaan Malaysia	Malásia
Abdulmajid Mohamed	4	Sebha University	Líbia
Workneh Y. Ayele	3	Stockholm University	Suécia
Shyamala Doraisamy	2	Universiti Putra Malaysia	Malásia
Kasper Christensen	2	Norwegian University of Life Sciences	Noruega
Sladjana Nørskov	2	Aarhus University	Dinamarca
Joachim Scholderer	2	Aarhus University	Dinamarca
Alwyn Vwen Yen Lee	2	Nanyang Technological University	Singapura
Seng Chee Tan	2	Nanyang Technological University	Singapura
Marina Carradore Sérgio	2	Universidade Federal de Santa Catarina	Brasil

Alexandre L. Gonçalves	2	Universidade Federal de Santa Catarina	Brasil
Gustaf Juell-Skielse	2	Stockholm University	Suécia
Mari-Sanna Paukkeri	1	Helsinki University of Technology	Finlândia
Tanja Kotro	1	National Consumer Research Centre	Finlândia
Anita Prinzie	1	The University of Manchester	Reino Unido
Sarah Herberz	1	Fraunhofer INT	Alemanha
Xingsen Li	1	Ningbo Institute of Technology	República da China
Liping Li	1	Computer and Information Institute	República da China
Zhengxin Chen	1	College of Information Science & Technology	Estados Unidos
Mark Klein	1	Massachusetts Institute of Technology	Estados Unidos
Ana Cristina B. Garcia	1	Universidade Federal Fluminense	Brasil
Haixia Liu	1	University of Nottingham Malaysia	Malásia
James Goulding	1	University of Nottingham	Reino Unido
Tim Brailsford	1	University of Nottingham Malaysia	Malásia
Amiya Tripathy	1	Don Bosco Institute of Technology	Índia
Suman Raina	1	Don Bosco Institute of Technology	Índia
Rylan Mascarenhas	1	Don Bosco Institute of Technology	Índia
Sonu Pangotra	1	Don Bosco Institute of Technology	Índia
Rilesh Rodrigues	1	Don Bosco Institute of Technology	Índia
Lars Frederiksen	1	Aarhus University	Dinamarca
Kristian Hovde Liland	1	Norwegian University of Life Sciences	Noruega
Knut Kvaal	1	Norwegian University of Life Sciences	Noruega
Einar Risvik	1	Nofima A/S	Noruega
Alessandra Biancolillo	1	Nofima A/S	Noruega
Tormod Næs	1	Nofima A/S	Noruega
João Artur de Souza	1	Universidade Federal de Santa Catarina	Brasil
Jieun Kim	1	Massachusetts Institute of Technology	Estados Unidos
Yongtae Park	1	Massachusetts Institute of Technology	Estados Unidos
Anna T. Röltgen	1	Abteilung für Arbeits Universität Trier	Alemanha
Valeria Bernardy	1	Abteilung für Arbeits Universität Trier	Alemanha
Rebecca Müller	1	Abteilung für Arbeits Universität Trier	Alemanha
ConnyH.Antoni	1	Abteilung für Arbeits Universität Trier	Alemanha
Sohee Ha	1	Seoul National University of Science and Technology	República da Coreia
Youngjung Geum	1	Seoul National University of Science and Technology	República da Coreia
Sercan Ozcan	1	University of Portsmouth	Reino Unido
Metin Suloglu	1	Leeds University	Reino Unido
C. Okan Sakar	1	Bahcesehir University	Turquia
Sushant Chatufale	1	Aston University	Reino Unido

Fonte: autor

E, por fim, o Quadro 2 apresenta as principais informações dos artigos, os quais foram lidos na íntegra.

Quadro 2 - Principais informações dos artigos da revisão sistemática

Autor	**Título**	**Ano**	**Palavras-Chave**	**Base de dados**	**Descrição Geral**
Paukkeri, M.-S., Kotro, T.	*Framework for analyzing and clustering short message database of ideas*	2009	*idea tool, innovation, text mining, practice, organizational memory*	Scopus	Método para salvar e organizar ideias em organizações
Thorleuchter, D., Van Den Poel, D., Prinzie, A.	*Mining ideas from textual information*	2010	*Idea mining, Text mining, Text classification, Technology*	Scopus	Método computacional de mineração de ideias
Thorleuchter, D., Herberz, S., Van Den Poel, D.	*Mining social behavior ideas of Przewalski horses*	2011	*Idea Mining, Social Behavior, Przewalski Horses, Textmining, Knowledge Discovery*	Scopus	Método de mineração de ideias para domínio social
Thorleuchter, D., Van Den Poel, D.	*Extraction of ideas from microsystems technology*	2012	*Idea Mining, Microsystems, Technology, Textmining, Knowledge Discovery*	Scopus	Método de mineração de ideias para domínio tecnológico
A. Tripathy; S. Raina; R. Mascarenhas; S. Pangotra; R. Rodrigues	*Extracting new product ideas from consumer blogs*	2012	*Web Logs;Data Mining;Search Engines;Stemming;Knowledge Discovery*	IEEE	Método de mineração de Web
Thorleuchter, D., Van Den Poel, D.	*Web mining based extraction of problem solution ideas*	2013	*Web mining, R&D planning, Idea mining, Text mining*	Scopus	Método computacional de mineração de ideias voltado para Web
Xingsen LiLiping LiZhengxin Chen	*Toward Extenics-Based Innovation Model on Intelligent Knowledge Management*	2014	*Extension innovation · Intelligent knowledge management ·Data mining · Extenics*	Springer Link	Modelo que combina Extenics, mineração de dados e gestão do conhecimento

Mark Klein, Ana Cristina Bicharra Garcia,	*High-speed idea filtering with the bag of lemons,*	2015	*Keywords: Collective intelligence; Open innovation; Social computing; Idea filtering*	Science Direct	Método para classificar ideias utilizando especialistas
Liu, H., Goulding, J., Brailsford, T.	*Towards computation of novel ideas from corpora of scientific text*	2015	*Idea mining · Text mining · Natural language processing · Recommender systems · Collaborative filtering*	Scopus	Método computacional de mineração de ideias
Thorleuchter, D., Van den Poel, D.	*Idea mining for web-based weak signal detection*	2015	*Web mining, Strategic decision making, Idea mining, Weak signal analysis*	Scopus	Método que usa mineração de ideias para detectar informação relevante (sinais fortes)
Alksher, M.A., Azman, A., Yaakob, R., Kadir, R.A., Mohamed, A., Alshari, E.M.	*A review of methods for mining idea from text*	2016	*Text mining; Idea mining; Information retrieval; Text classification*	Scopus	Revisão geral de métodos de mineração de ideias
Thorleuchter, D., Van den Poel, D.	*Identification of interdisciplinary ideas*	2016	*Text mining, Latent semantic indexing, Idea mining, Clustering, Classification*	Scopus	Método de mineração de ideias interdisciplinares usando cluster e classificação semântica
Kasper Christensen, Sladjana Nørskov, Lars Frederiksen and Joachim Scholderer	*In Search of New Product Ideas: Identifying Ideas in Online Communities byMachine Learning and Text Mining*	2017	*no*	NA	Mineração de ideias utilizando machine learning
Alksher, M.A., Azman, A., Yaakob, R., Abdul Kadir, R., Mohamed, A., Alshari, E.	*A framework for idea mining evaluation*	2017	*Text mining, Idea mining, Evaluation Process, Idea Reliability*	Scopus	Estrutura baseada em análise estatística para avaliação humana dos resultados da mineração de ideias

Kasper Christensen, Kristian Hovde Liland, Knut Kvaal, Einar Risvik, Alessandra Biancolillo, Joachim Scholderer, Sladjana Nørskov, Tormod Næs	*Mining online community data: The nature of ideas in online communities*	2017	*Machine learning, Text mining, Natural language processing, Online communities, ideas, Partial least squares, Support vector machines*	Science Direct	Método de mineração de Web
Alwyn Vwen Yen Lee & Seng Chee Tan	*Discovering dynamics of an idea pipeline: Understanding idea development within a knowledge building discourse*	2017	*Online discourse analysis, knowledge building discourse, idea pipeline, idea development, text mining, network analysis*	NA	Framework de dinâmica do desenvolvimento de ideias dentro de uma construção de conhecimento
Alwyn Vwen Yen Lee and Seng Chee Tan	*Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis*	2017	*Temporal analytics, machine learning, cluster analysis, promising ideas, idea analysis, knowledge building discourse*	NA	Método de mineração usando análise temporal e análise de cluster
M. Carradore Sergio; J. Artur de Souza; A. Leopoldo Goncalves	*Idea Identification Model to Support Decision Making*	2017	*Cluster Analysis;Idea Management;Innovation;Ontology*	IEEE	modelo baseado em ontologias e análise de cluster para apoiar a Gestão de Ideias
Mostafa Alksher, Azreen Azman, Razali Yaakob, Rabiah Abdul Kadir, Abdulmajid Mohamed and Eissa Alshari	*Feasibility of Using the Position as Feature for Idea Identification from Text*	2018	*Text extraction; Text pattern; Text position; Idea identification; Information retrieval*	NA	Método para identificar e posicionar ideias dentro de artigos
Alksher M., Azman A., Yaakob R., Alshari E.M., Kadir R.A., Mohamed A.	*Effective idea mining technique based on modeling lexical semantic*	2018	*Idea mining, Information retrieval, WordNet, text pattern, text mining*	Scopus	Método baseado em semântica para melhorar a relação entre termos de mineração de ideias

Marina Carradore Sérgio; Alexandre Leopoldo Gonçalves	*Análise e interpretação de ideias: proposta de um modelo*	2019	*Idea management; Idea mining; Innovation; Idea mining model*	Scopus	Modelo de análise e interpretação de ideias
Azreen Azman, Mostafa Alksher, Eissa Alshari, Razali Yaakob, Shyamala Doraisamy	*Optimization of idea mining model based on text position weight*	2019	*Idea mining, Text position, Optimization, Curve fitting, Artificial Neural Network*	Scopus	Otimização de modelo de mineração de ideias
Jieun Kim, Yongtae Park	*Leveraging ideas from user innovation communities: using text-mining and case-based reasoning*	2019	*no*	Scopus	Aplicativo de mineração de ideias com text mining e raciocínio baseado em casos
Azreen Azman, Mostafa Alksher, Shyamala Doraisamy, Razali Yaakob, and Eissa Alshari	*A Framework for Automatic Analysis of Essays Based on Idea Mining*	2020	*Online learning, Essay assessment, Idea mining, Text mining, Analysis tool*	Springer Link	Framework de mineração de ideias
Röltgen, A.T.Email Author, Bernardy, V., Müller, R., Antoni, C.H.	*Development, implementation and evaluation of a digital idea management system. A case analysis*	2020	*Digital collaboration, Formative evaluation, Idea management, Innovation*	Scopus	Criação de aplicativo de gestão de ideias
Workneh Y. Ayele	*A Data Mining Process for Generating Ideas Using a Textual Dataset*	2020	*CRISP-IM; idea generation; idea evaluation; idea mining evaluation; dynamic topic modeling; CRISP-DM*	Springer Link	Aplicativo de mineração de ideias
Workneh Y. Ayele; Gustaf Juell-Skielse	*A Process Model for Generating andEvaluating Ideas: The Use of MachineLearning and Visual Analytics to SupportIdea Mining*	2020	*Idea mining Idea generation Idea evaluation Text mining Machine learning Dynamic topic modeling*	Springer Link	Aplicativo de mineração de ideias, melhoria

Ayele W.Y., Juell-Skielse G	*A Systematic Literature Review about Idea Mining : The Use of Machine-Driven Analytics to GenerateIdeas*	2021	*Computer-assisted creativity, Idea elicitation, Idea generation, Idea mining, Machinelearning , Machine-driven analytics, Text mining*	Scopus	Revisão sistemática de métodos de mineração de ideias
Sercan Ozcan, Metin Suloglu, C. Okan Sakar, Sushant Chatufale	*Social media mining for ideation: Identification of sustainable solutions and opinions*	2021	*Text mining Semi-supervised learning Support vector machines Decision-making Crowdsourcing Sustainability*	Science Direct	Modelo de classificação para identificar tweets que continha uma ideia
Sohee Ha, Youngjung Geum	*Identifying new innovative services using M&A data: An integrated approach of data-driven morphological analysis*	2022	*New service development Big data M&A QFD MA Data analytics*	Science Direct	Processo para a construção de implantação de função de qualidade orientada a dados (QFD) e análise morfológica baseada em dados (MA)

Fonte: autor

Com a finalidade de permitir a organização dos 30 artigos analisados nesta revisão sistemática da literatura, estes foram divididos em grupos, conforme a técnica que foi aplicada para implementação da tarefa de mineração de ideias. Os grupos propostos foram: artigos de revisão da literatura, artigos com métodos manuais de mineração de ideias, artigos com métodos automáticos baseados em regras e artigos com métodos automáticos baseados em aprendizado de máquina. O Quadro 3 apresenta os dados acima descritos.

Quadro 3 - Técnica aplicada na tarefa de mineração de ideias

Técnica de Mineração de Ideias Aplicada	Número de artigos	Referência do Artigo
Artigos sem técnica de Mineração de Ideias (Revisão da Literatura)	2	(ALKSHER et al., 2016; AYELE; JUELL-SKIELSE, 2021)
Artigos com Métodos Manuais de Mineração de Ideias	2	(KLEIN; GARCIA, 2015; PAUKKERI, 2009)
Artigos com Métodos Automáticos de Mineração de Ideias Baseados em Regras	14	(ALKSHER et al., 2017, 2018b, 2018a; KIM; PARK, 2019; LEE; TAN, 2017b; LI; LI; CHEN, 2014; LIU; GOULDING; BRAILSFORD, 2015; THORLEUCHTER; HERBERZ; POEL, 2011; THORLEUCHTER; VAN DEN POEL, 2015, 2016, 2012, 2013; THORLEUCHTER; VAN DEN POEL; PRINZIE, 2010; TRIPATHY et al., 2012)
Artigos com Métodos Automáticos de Mineração de Ideias Baseados em Aprendizado de Máquina	12	(AYELE, 2020; AYELE; JUELL-SKIELSE, 2020; AZMAN et al., 2019, 2020; CHRISTENSEN et al., 2017a, 2017b; HA; GEUM, 2022; LEE; TAN, 2017a; OZCAN et al., 2021; RÖLTGEN et al., 2020; SÉRGIO; GONÇALVES, 2019; SÉRGIO; SOUZA; GONÇALVES, 2017)

Fonte: autor

Pode-se verificar através desta organização realizada que há uma tendência de utilização de métodos automáticos de mineração de ideias baseados em aprendizado de máquina nos últimos 5 anos. Este parece ser um campo promissor que ainda pode ser aperfeiçoado e que possui vantagens e desvantagens se comparado com técnicas mais tradicionais. Pelo fato de utilizarem aprendizado supervisionado, o conceito de ideia torna-se simplificado, visto que será apresentado ao sistema exemplos de ideias selecionadas por especialistas dentro de um conjunto de dados. Porém, essa seleção de ideias pode ser subjetiva e tornar o método pouco efetivo, uma vez que o sucesso da aprendizagem de máquina está altamente relacionado à qualidade dos dados.

3.4 DESENVOLVIMENTO DA PESQUISA

Esta pesquisa teve seu desenvolvimento seguindo a abordagem metodológica da *Design Science Research*, que tem como objetivo prover um processo para a condução de pesquisas científicas e fornecer aos pesquisadores um modelo para detalhar e apresentar resultados de pesquisa (PEFFERS *et al*., 2007).

O capítulo inicial da tese, mais especificamente as seções 1.2 e 1.5, assim como a fundamentação teórica contêm o detalhamento do que foi realizado na etapa 1. Ainda neste capítulo, mais especificamente na seção 1.4, está contida a definição dos objetivos apontados pela etapa 2 da DSR. No que tange ao estado da arte, mencionadas nas etapas 1 e 2, foi realizada uma revisão sistemática da literatura, detalhada na seção 3.3.

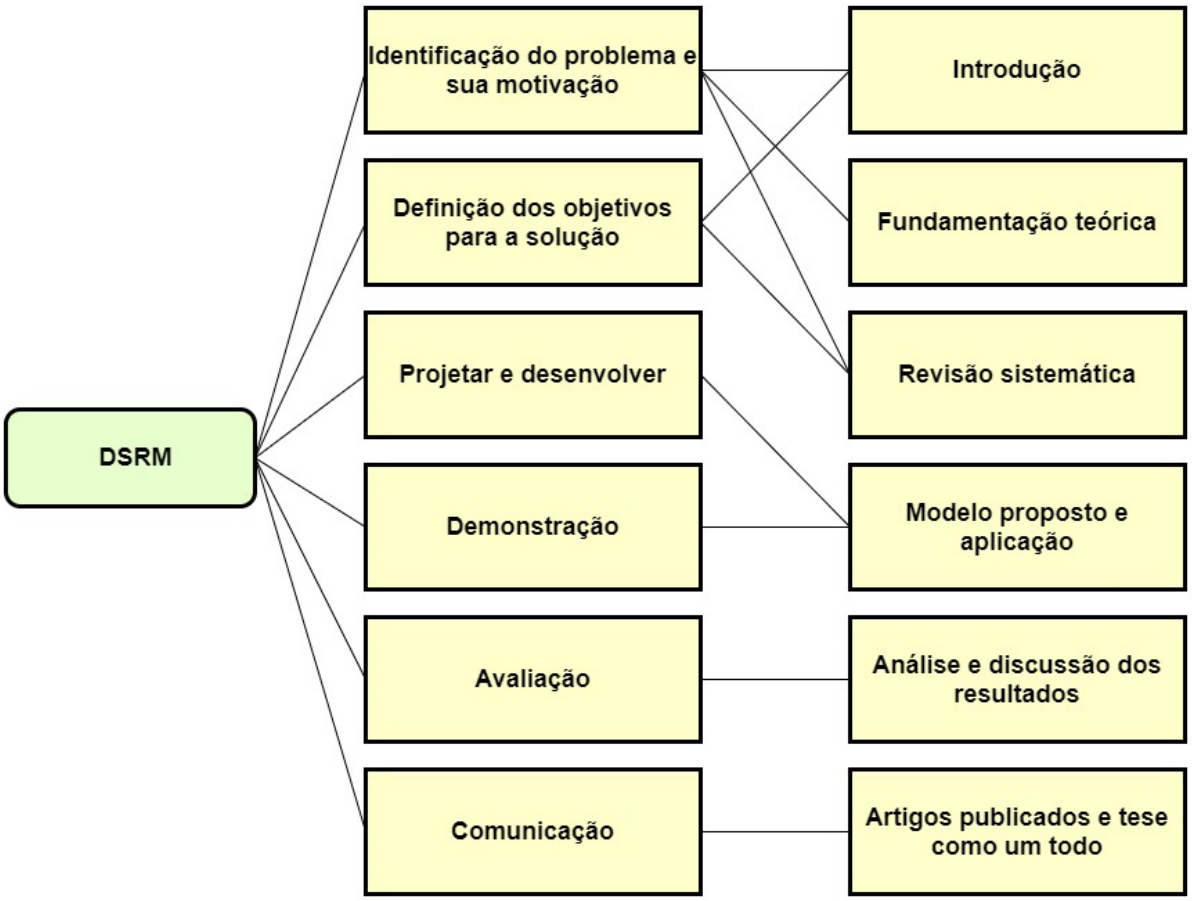
No capítulo 4 é apresentado como o artefato, ou seja, o modelo proposto, foi construído, a partir de testes realizados levando em consideração as etapas 3 e 4. Neste capítulo, todas as etapas do modelo e a interconexão entre as mesmas são detalhadas. Ainda, para tornar mais claro o entendimento da proposição realizou-se uma instanciação do modelo.

O capítulo 5 contempla a etapa 5. Após a demonstração dos resultados, estes foram analisados e discutidos, permitindo uma avaliação do modelo proposto.

E, por fim, a etapa de comunicação foi realizada através de publicação de artigos em periódicos e eventos, bem como a própria escrita da tese como um todo.

Os relacionamentos das etapas da DSRM com as seções da tese correspondentes são mostrados na Figura 7.

Figura 7 – Seções da tese e etapas da DSRM



Fonte: autor

As seções a seguir promovem um detalhamento das etapas do desenvolvimento da tese, segundo as etapas da DSRM.

3.4.1 Identificação do problema e sua motivação

A identificação do problema foi feita a partir da revisão sistemática, detalhada na seção 3.3. O problema e sua motivação também são descritos na seção 1.2.

3.4.2 Definição dos objetivos para a solução

Na sequência, os objetivos desta tese foram definidos levando-se em conta o problema identificado a partir da revisão sistemática (seção 3.3) e estão descritos na seção 1.4.

3.4.3 Projetar e Desenvolver

Para projetar e desenvolver o sistema utilizado na avaliação do modelo foram utilizados diferentes ambientes de desenvolvimento, descritos nos parágrafos a seguir.

Para a classificação dos conjuntos de dados utilizou-se a aplicação *Lightside*® (MAYFIELD; ROSÉ, 2012), por ser um aplicativo de código aberto destinado à tarefa de classificação de textos. Foram utilizadas as configurações padrões do classificador SVM, habilitando a opção *LibLinear*, adequada para aumentar a eficiência do classificador. Também foram utilizadas as configurações padrões do classificador DT, desabilitando as opções de trabalho com tabelas esparsas e de alta dimensão, condição mais indicada para trabalhos que envolvem valores numéricos. E por último, as configurações padrões do classificador NB foram utilizadas, onde se encontram desabilitadas as opções de utilização de estimadores *Kernel* e discretização supervisionada.

Para a configuração dos recursos foi utilizada a opção de *unigrams* onde são extraídas as palavras distintas do conteúdo de cada texto do conjunto de dados. Cada palavra representa uma característica, tendo agregada a si a frequência em que ocorre na coleção de textos. Também foi utilizada a opção *track feature hit location* que inclui o registro da localização de cada característica extraída de cada documento. Isso é importante para permitir que se execute a análise de erros após a criação e avaliação de um modelo (MAYFIELD; ROSÉ, 2012).

Como a aplicação *Lightside*® não contempla o classificador RF, a aplicação escolhida neste caso foi o Orange® (DEMSAR *et al.*, 2016) por ser um *kit* de ferramentas de análise e visualização de dados de código aberto, aprendizado de máquina e mineração de dados. Foram utilizadas as configurações padrões do classificador RF, utilizando *replicable training* (treinamento replicável) e desabilitando a opção *growth control* (controle de crescimento).

Para a composição dos KGs do conjunto de dados utilizou-se a aplicação *Sobek Mining*® (REATEGUI *et al.*, 2011), por ser um aplicativo de mineração de texto que gera um KG como resultado. A *Sobek Mining*® é uma ferramenta de mineração de texto criada em 2007 pelo Grupo de Pesquisa em Tecnologia Aplicada à Educação do Programa de Pós-Graduação em Informática na Educação da Universidade Federal do Rio Grande do Sul. Sua aplicação está destinada à tarefas que envolvam a compreensão da leitura e criação de resumos, em que através da mineração de texto, cria um grafo do texto a partir dos conceitos mais relevantes e suas relações utilizando como princípio a análise de frequência de cada

termo. A aplicação permite a escolha de criação de grafos com 15, 30 ou 50 conceitos, sendo que para esta tese se utilizou a configuração de 50 conceitos por ser a configuração máxima, com a finalidade de captar o maior número possível de conceitos.

O método de composição dos grafos de conhecimento usado pelo *Sobek Mining*® foi construído baseado no modelo de gráfo de distância simples, em que os nós representam os principais termos encontrados no texto e as arestas usadas para ligar os nós representam informações de adjacência. Portanto, nós e arestas representam como os termos aparecem no texto. O método também identifica termos compostos e, par tal, se utiliza de alguns parâmetros para extrair os conceitos compostos com mais de uma palavra: frequência mínima (indica o menor número de ocorrências que uma palavra deve ter para aparecer no gráfico), lista de cada conceito e seus vizinhos, número de ocorrências de cada conceito e número máximo de conexões possíveis. De acordo com estes parâmetros, é criada uma combinação da palavra atual com as palavras subsequentes. Depois de identificar as combinações frequentes de palavras, chamadas de conceitos, o processo de mineração seleciona o conjunto de conceitos relevantes com base em sua frequência no texto. A próxima etapa calcula a semelhança entre conceitos. O coeficiente de similaridade entre eles são calculados com o produto escalar. Também é calculado o coeficiente de relevância para comparar dois conceitos e manter aquele de maior importância, mesmo que este possua uma frequência menor (REATEGUI et al., 2011).

Para o cálculo de coordenadas de WEs foi utilizado o método *Word2vec*® (MIKOLOV et al., 2013) por ser uma técnica para processamento de linguagem natural que utiliza um modelo de rede neural para aprender associações de palavras a partir de um grande corpo de texto. O modelo de aprendizagem utilizado foi o CBOW, com 50 dimensões e no idioma português. Os algoritmos criados foram implementados em *Python*® (VAN ROSSUM, 1995) .

Nas seções a seguir serão detalhadas as etapas de pré-processamento (comum aos classificadores, KGs e WEs), transformação vetorial (utilizada nos classificadores) e treinamento (comum aos classificadores, KGs e WEs).

3.4.3.1 Pré-Processamento

O pré-processamento é uma etapa necessária para a retirada de informações do texto que não são relevantes ao contexto de estudo. Para isso foram realizadas a tokenização e a remoção de *stopwords*.

A tokenização é o processo de transformação de um texto em unidades menores, chamados de *tokens*, responsáveis por realizar a segmentação do texto. Quando estes *tokens* isolados de seus contextos não exprimem significados determinantes na compreensão do texto, ocorre a sua eliminação através do processo denominado de remoção das *stopwords* (JURAFSKY; MARTIN, 2020).

3.4.3.2 Transformação Vetorial

Para a correta utilização de cada um dos documentos que compõem a coleção de documentos pelos classificadores torna-se necessária a realização de uma transformação vetorial. Neste trabalho, foi utilizado o método *bag-of-words*.

O método *bag-of-words* representa um vetor de documento constituído de palavras (ou características) e suas frequências ou pesos e, após a sua constituição, serve de entrada para a classificação de documentos em que a frequência de ocorrência ou peso de cada característica é usada como seu recurso para treinar o classificador de texto (HARRIS, 1954; DHAVALIKAR; CHOUDHARI, 2021).

3.4.3.3 Treinamento

O procedimento de validação cruzada foi aplicado a todos os MTECs deste trabalho que envolvam treinamento. Para o caso específico dos classificadores, além da validação cruzada foi utilizada a separação do conjunto de dados em 80% treinamento e 20% testes.

A validação cruzada é um procedimento estatístico bastante comum em atividades de mineração de dados e aprendizado de máquina, utilizado basicamente para evitar resultados tendenciosos ao se utilizar um espaço amostral reduzido. Consiste na partição de uma amostra de dados em subconjuntos mutuamente exclusivos onde a análise é inicialmente realizada em um subconjunto enquanto os demais são guardados para uma confirmação e validação da análise inicial. O conjunto de dados inicial é chamado de conjunto de treinamento e os outros são chamados de conjunto de validação ou teste. O procedimento é repetido circularmente

para todas as partições, onde cada uma é utilizada uma única vez durante o treinamento. O resultado final é a média das avaliações (KOHAVI, 1995; SCHAFFER, 1993; TAVARES; LOPES; LIMA, 2007).

3.4.3.4 Cálculo do delimitador para os Word Embeddings

Para o treinamento dos WEs foi necessário a criação de um delimitador entre as ideias e os textos comuns, utilizando-se do centroide das 5 palavras mais frequentes do texto analisado, aplicando a seguinte equação:

$$centroide_{coord.} = minimo_{coord.} + \left(\frac{maximo_{coord.} - minimo_{coord.}}{2}\right) \quad (1)$$

A partir do centroide geral dos textos das ideias traçou-se uma figura geométrica delimitadora que pudesse otimizar esta separação entre os textos contendo ideias e textos comuns. Utilizando o centroide geral dos textos das ideias como ponto central, a figura foi elaborada variando suas distâncias delimitadoras dos eixos *x, y* e *z*. Os valores foram obtidos após testes com diferentes configurações, apresentados na Tabela 1.

Tabela 1 - Diferentes configurações para a obtenção da figura delimitadora com as respectivas acurácias

Identificador do teste	Distância somada ao centroide geral dos textos das ideias no eixo X	Distância somada ao centroide geral dos textos das ideias no eixo Y	Distância somada ao centroide geral dos textos das ideias no eixo Z	Acurácia
1	0,1	0,1	0,1	59%
2	0,2	0,2	0,2	68%
3	0,3	0,3	0,3	65%
4	0,4	0,4	0,4	56%
5	0,2	0,3	0,1	71%
6	0,3	0,2	0,1	69%
7	0,25	0,26	0,10	73%

Fonte: autor

As distâncias foram calculadas inicialmente utilizando variações de 0.1 a partir do centroide. Verificou-se então que a partir da distância 0.5, a acurácia dos testes caiu consideravelmente. Portanto, as variações de 0.1 foram realizadas até 0.4, onde foram testadas todas as possibilidades, porém, apenas apresentadas acima são as mais relevantes. Entre as

configurações 5 e 6, foram obtidos os melhores desempenhos. Visando aferir a possibilidade de incremento na acurácia, foram aplicadas variações de passo na segunda casa decimal de 0.01 nas distâncias de *x*, *y* e *z* para os valores que constam entre as configurações 5 e 6.

O melhor resultado está identificado como 7 e possibilitou a obtenção da matriz confusão apresentada na Tabela 2.

Tabela 2 - Matriz confusão entre ideias e textos comuns considerando os resultados dos WEs

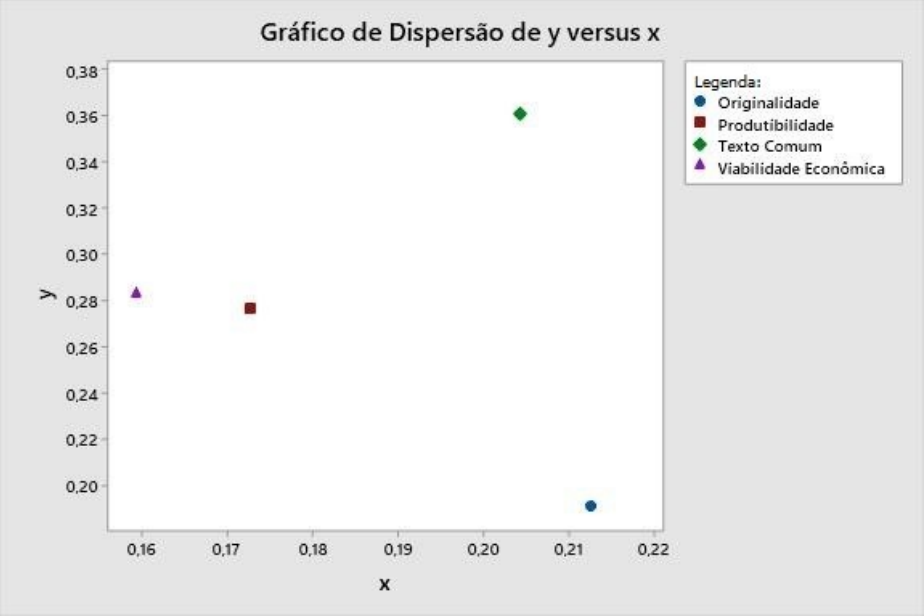
Atual / Predito	ideia	texto comum
ideia	86	36
texto comum	26	74

Fonte: autor

A matriz confusão mostra que o índice de acerto dos textos contendo ideias foi de 71% e dos textos comuns de 74%, atingindo uma acurácia total em torno de 73%.

Os mesmos procedimentos descritos foram realizados considerando os WEs treinados para identificar os critérios dos especialistas versus textos comuns. A Figura 8 apresenta o gráfico de dispersão em 2 dimensões para facilitar a visualização dos centroides gerais das ideias, das ideias separadas por critérios e dos textos comuns.

Figura 8 – Gráfico dos centroides



Fonte: autor

Nesta figura é possível verificar que existe separabilidade linear entre os critérios e os textos comuns, mostrando que os WEs são capazes de realizar a tarefa de identificação dos critérios.

E, por fim, é realizada a comparação entre os centroides gerais de cada critério com cada texto do conjunto de dados, sendo apresentada abaixo a matriz confusão na Tabela 3.

Tabela 3 - Matriz confusão entre critérios e textos comuns considerando os resultados dos WEs

Atual / Predito	ideia por critério	texto comum
ideia por critério	113	9
texto comum	24	76

Fonte: autor

Como se trata de WEs individuais, para contabilização dos resultados foi realizado o mesmo procedimento descrito para os classificadores nesta condição, onde são considerados os acertos dos WEs individuais treinados para identificar cada critério. O acerto final do WE é atingido se os resultados obtidos alcancem a condição de acerto de pelo menos um dos critérios.

3.4.3.5 Definição das Regras de Comparação dos Knowledge Graphs

Após a submissão dos textos separados por critérios ao KG e encontrados os termos mais frequentes e as relações que remetem ao critério, foram criadas regras de aderência para cada critério específico. O algoritmo que define as regras de aderência compara o KG gerado para cada critério com um grafo simples produzido por um texto no *Sobek Mining*®. Quanto maior o número de conceitos e relações em comum, maior a possibilidade de um texto ser aderente a um critério. Foram considerados no mínimo 3 conceitos em comum e pelo menos 1 relação em comum, como valores de corte.

3.4.3.6 Cálculo do Ranking

Para realização dos cálculos do *ranking* e do grau de pertinência aos critérios foram definidas as seguintes variáveis, declaradas no Quadro 4.

Quadro 4 – Definição das variáveis para cálculo do ranking

Variável	Descrição
C_O	Resultados do classificador treinado para identificação do critério Originalidade
C_P	Resultados do classificador treinado para identificação do critério Produtibilidade
C_V	Resultados do classificador treinado para identificação do critério Viabilidade Econômica
KG_O	Resultados do KG treinado para identificação do critério Originalidade
KG_P	Resultados do KG treinado para identificação do critério Produtibilidade
KG_V	Resultados do KG treinado para identificação do critério Viabilidade Econômica
WE_O	Resultados dos WE treinado para identificação do critério Originalidade
WE_P	Resultados dos WE treinado para identificação do critério Produtibilidade
WE_V	Resultados dos *WE* treinado para identificação do critério Viabilidade Econômica
Res_ C	Composição dos resultados dos 3 critérios avaliados pelo classificador, que indica se pelo menos um critério foi identificado por um classificador treinado para cada critério
Res_KG	Composição dos resultados dos 3 critérios avaliados pelo *Knowledge Graph*, que indica se pelo menos um critério foi identificado por um KG treinado para cada critério
Res_WE	Composição dos resultados dos 3 critérios avaliados pelo *Word Embedding*, que indica se pelo menos um critério foi identificado por um WE treinado para cada critério
G(O)	Grau de pertinência ao critério Originalidade
G(P)	Grau de pertinência ao critério Produtibilidade
G(V)	Grau de pertinência ao critério Viabilidade Econômica

Fonte: autor

Vale mencionar que diferentes classificadores podem ser utilizados, sendo necessário, para determinada execução, a escolha de um classificador em particular. A seguir é apresentada a equação que compões o resultado do *ranking*:

$$ranking = 100 \times \left(\frac{Res\_C+Res\_KG+Res\_WE+C\_0+KG\_O+WE\_O+C\_P+KG\_P+WE\_P+C\_V+KG\_V+WE\_V}{12}\right) \quad (2)$$

As variáveis Res_C , Res_KG e Res_WE (composição dos resultados do classificador, do KG e do WE) indicam se um texto é ou não uma ideia. E as variáveis *C_O*, *C_P*, *C_V*, *KG_O*, *KG_P*, *KG_V*, *WE_O*, *WE_P* e *WE_V*, (resultados individuais do classificador, do KG e do WE) indicam os acertos de cada critério. Na equação do *ranking* é feito uma média entre a composição dos resultados do classificador, do KG e do WE e os resultados individuais de cada critério para o classificador, o KG e o WE.

Os graus de pertinência de cada um dos critérios são calculadas da seguinte forma:

$$G(O) = \left(\frac{(C\_O+KG\_O+WE\_O)}{(C\_O+KG\_O+WE\_O)+(C\_P+KG\_P+WE\_P)+(C\_V+KG\_V+WE\_V)}\right) \times 100 \quad (3)$$

$$G(P) = \left(\frac{(C\_P+KG\_P+WE\_P)}{(C\_O+KG\_O+WE\_O)+(C\_P+KG\_P+WE\_P)+(C\_V+KG\_V+WE\_V)}\right) \times 100 \quad (4)$$

$$G(V) = \left(\frac{(C\_V+KG\_V+WE\_V)}{(C\_O+KG\_O+WE\_O)+(C\_P+KG\_P+WE\_P)+(C\_V+KG\_V+WE\_V)}\right) \times 100 \quad (5)$$

3.4.4 Demonstração

A demonstração da viabilidade do modelo foi construída a partir de cenários de estudo. Os conjuntos de dados utilizados para elaboração destes cenários foram divididos em conjunto de dados para avaliação do modelo e conjunto de dados para verificação do modelo. O objetivo desta divisão ocorreu pois, para avaliação do modelo foi necessário um conjunto de dados voltado para avaliação das técnicas, onde as ideias e os textos comuns estivessem claramente separados. Isto se justifica devido a escolha dos métodos para cada tarefa do modelo sendo interessante um treinamento com as ideias claramente evidenciadas em relação aos textos comuns. Porém, em uma condição prática e real, isso dificilmente ocorria. Sendo assim, foi criado um conjunto de dados para verificação do modelo onde os textos que não contêm ideias são mais semelhantes aos textos que contém ideias, porém sem indicativos de inovação.

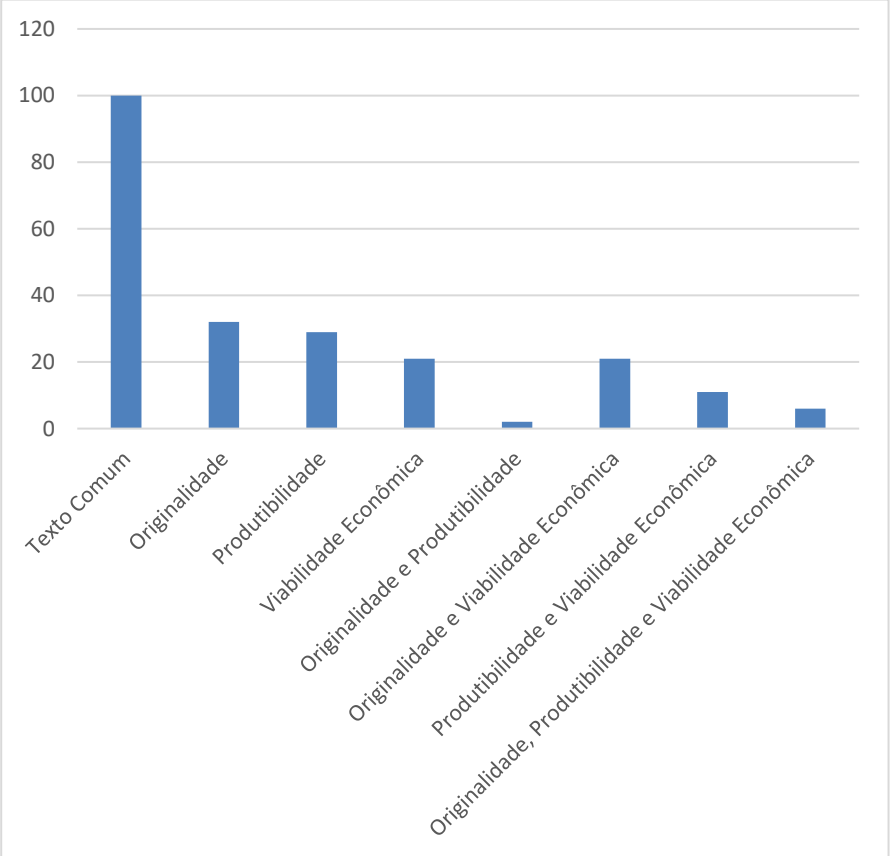
3.4.4.1 Conjunto de Dados para Avaliação Inicial do Modelo

Este conjunto de dados contém ideias disponibilizadas publicamente através do Portal Sinapse da Inovação®, que é um programa de incentivo ao empreendedorismo inovador que tem por objetivo “transformar e aplicar as boas ideias geradas por estudantes, pesquisadores, professores e profissionais dos diferentes setores do conhecimento e econômicos em negócios de sucesso” (SINAPSE DA INOVAÇÃO, 2018). O conjunto de dados possui 122 textos representando ideias que alcançaram a última etapa para serem selecionadas, sendo aprovadas e que receberam aporte financeiro do Sinapse na Inovação®,

edição 2018. Adicionalmente, foram incluídos ao conjunto de dados 100 textos com conteúdo diverso retirado da *web*.

O conjunto de dados de ideias foi então classificado manualmente considerando os três critérios de especialistas considerados nas avaliações iniciais, sendo originalidade, produtibilidade e viabilidade econômica, a partir da leitura individual de cada ideia. Os quantitativos foram: ideias que explicitamente mencionam somente a viabilidade econômica (rótulo Viabilidade Econômica), 21; ideias que mencionam somente informações que possam representar o critério de produtibilidade (rótulo Produtibilidade), 29 e ideias que em seu texto denotam um grau elevado somente do critério de originalidade (rótulo Originalidade), 32. Existem textos em que aparecem mais de um critério, sendo estes rotulados de Originalidade e Produtibilidade, 2, Originalidade e Viabilidade Econômica, 21, Produtibilidade e Viabilidade Econômica, 11 e Originalidade, Produtibilidade e Viabilidade Econômica, 6. Por fim, foram rotulados os 100 textos comuns extraídos aleatoriamente da Wikipedia® (rótulo Texto Comum). Os quantitativos são apresentados na Figura 9.

Figura 9 - Separação do conjunto dados para avaliação do modelo (textos comuns x ideias classificadas por critérios)



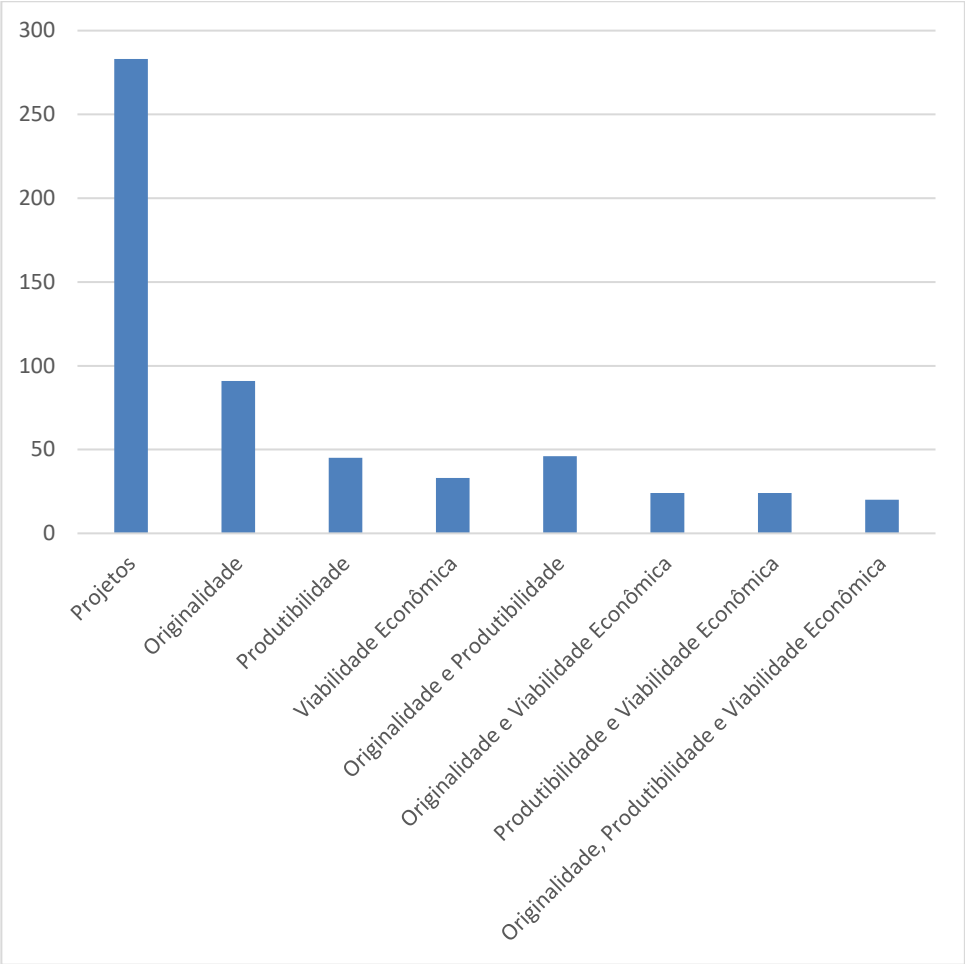
Fonte: Autor

3.4.4.2 Conjunto de Dados para Verificação do Modelo

Este conjunto de dados é composto por 283 ideias retiradas da lista das iniciativas premiadas pelo Concurso Inovação da Escola Nacional de Administração Pública – ENAP®. Para a verificação, o conjunto de dados foi lido e classificado manualmente para extrair os critérios utilizados na seleção de cada ideia. De maneira complementar, foi adicionado outro conjunto de dados com 283 textos contendo projetos aprovados pela Universidade Federal do Rio Grande - FURG e cadastrados no Sistema de Controle de Projetos – SisProj. Mesmo sendo projetos aprovados, estes não representam ideias contendo inovação e por isto foram utilizados como contraponto aos textos contendo ideias. As 283 ideias e os 283 textos foram extraídos a partir do Portal de Dados Abertos® (BRASIL, 2012).

A leitura e classificação manual identificou 91 ideias que mencionam somente o critério originalidade, 45 ideias que mencionam somente o critério produtibilidade e 33 ideias que mencionam somente o critério viabilidade econômica. O conjunto também apresenta ideias contendo mais de um critério, sendo elas 46 ideias que mencionam os critérios originalidade e produtibilidade, 24 ideias que mencionam os critérios originalidade e viabilidade econômica, 24 ideias que mencionam os critérios produtibilidade e viabilidade econômica e, por fim, 20 ideias que mencionam os critérios originalidade, produtibilidade e viabilidade econômica. Os quantitativos são apresentados na Figura 10.

Figura 10 - Separação do conjunto dados para verificação do modelo (projetos x ideias classificadas por critérios)



Fonte: Autor

3.4.5 Avaliação

Para avaliação de desempenho dos classificadores de texto e para o resultado ordenação (*ranking*) das ideias, foram utilizadas as métricas de acurácia, precisão (*precision*), revocação (*recall*), *f1-score* e o coeficiente *kappa*. Estas métricas são baseadas na matriz confusão.

A matriz confusão, também conhecida como matriz de erro ou matriz de confusão, representa um *layout* de tabela específico que permite a visualização do desempenho de um algoritmo, tipicamente voltado ao aprendizado supervisionado, mostrando as classificações corretas e incorretas.

De acordo com Campbell e Wynne (2011), os erros estão presentes em qualquer tipo de classificação e a forma padronizada para reportar erros em locais específicos é a chamada matriz confusão. Esta matriz identifica não somente o erro global da classificação para cada categoria, mas também como se deram as confusões entre as categorias. No Quadro 5 é apresentada a matriz confusão.

Quadro 5 – Apresentação da matriz confusão

		Classificação Prevista	
		P	N
Classificação Atual	P	VP	FN
	N	FP	VN

Fonte: autor

No quadro acima e nas equações são apresentadas a seguir, *VP* representa a classificação correta da classe positivo, *VN* representa a classificação correta da classe negativo, *FP* representa o erro em que o modelo previu a classe positivo quando o valor real era a classe negativo e *FN* representa o erro em que o modelo previu a classe negativo quando o valor real era a classe positivo.

A acurácia é uma métrica simples que informa o desempenho de um classificador ao prever rótulos de classe a partir de seu conjunto de dados. A equação 6 ilustra o cálculo da acurácia.

$$acurácia = \frac{VP+VN}{VP+VN+FP+FN} \tag{6}$$

A precisão é a capacidade de evitar falsos positivos na segmentação do conjunto de dados e pode ser usada em uma situação em que os falsos positivos são considerados mais prejudiciais que os falsos negativos. A equação 7 ilustra o cálculo da precisão.

$$precisão = \frac{VP}{VP+FP} \tag{7}$$

O revocação representa a proporção entre as segmentações corretas e o total de segmentações realizadas. Essa métrica indica o quão boa foi a segmentação na escolha dos pontos corretos do conjunto e dados, e pode ser usada em uma situação onde os falsos negativos são considerados mais prejudiciais que os falsos positivos. A equação 8 ilustra o cálculo da revocação.

$$revocação = \frac{VP}{VP+FN} \tag{8}$$

O f1-*score* é simplesmente uma maneira de observar somente 1 métrica ao invés de duas (precisão e revocação) em alguma situação. É uma média harmônica entre as duas, que está muito mais próxima dos menores valores do que uma média aritmética simples. Ou seja, quando se tem um f1-*score* baixo, é um indicativo de que ou a precisão ou a revocação está baixa. A equação 9 ilustra o cálculo do f1-*score*.

$$f1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

Já o coeficiente *kappa* é comumente utilizado para avaliar a concordância entre as classificações de dois avaliadores em uma escala nominal (DE RAADT *et al.*, 2019). A equação 10 ilustra o cálculo do coeficiente *kappa*.

$$kappa = \frac{P(O)-P(E)}{1-P(E)} \tag{10}$$

Na equação 10, *P(O)* representa a proporção observada de concordâncias (soma das respostas concordantes dividida pelo total) e *P(E)* representa a proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total).

3.4.6 Comunicação

A comunicação dos resultados foi realizada através de publicações em periódicos e congressos, visando apresentar os resultados obtidos à comunidade científica.

3.5 SÍNTESE DA METODOLOGIA DE PESQUISA

O enquadramento metodológico foi realizado sendo esta classificada como paradigma funcionalista, modalidade tecnológica e de natureza aplicada. Em seguida são detalhadas todas as etapas da DSRM que serviram de base para a construção do modelo propriamente dito.

A revisão sistemática foi o procedimento adotado para identificar as pesquisas relevantes sobre o tema MI através de um método sistemático.

E finalizando este capítulo a metodologia de pesquisa foram descritas todas as etapas da tese, bem como sua relação com a DSRM. No Quadro 6 é apresentada uma síntese do desenvolvimento da pesquisa.

Quadro 6 – Síntese do desenvolvimento da pesquisa

Etapa	Descrição
Identificação do problema e sua motivação	Revisão sistemática da literatura.
Definição dos objetivos para a solução	Revisão sistemática da literatura.
Projetar e Desenvolver	Diferentes ferramentas/métodos utilizadas desenvolvimento. Ferramentas: Lightside®, Orange®, Sobek Mining® ; Método: Word2vec; e Linguagem de Programação: Python®.
	Pré-processamento: *tokenização* e a remoção de *stopwords*.
	Transformação vetorial: *bag-of-words.*
	Treinamento: validação cruzada e separação do segundo conjunto de dados em treinamento/teste.
Demonstração	Dois cenários de estudo: • Conjunto de dados para avaliação inicial do modelo: 122 ideias e 100 textos comuns. • Conjunto de dados para os verificação do modelo: 283 ideias e 283 projetos aprovados.
Avaliação	Métricas de acurácia, precisão (*precision*), revocação (*recall*), f1-*score* e o coeficiente *kappa*. Estas métricas são baseadas na matriz confusão.
Comunicação	Publicações dos resultados em periódicos e congressos.

Fonte: autor

4 MODELO PROPOSTO

Este capítulo visa apresentar o modelo proposto detalhando e exemplificando cada uma das etapas que o compõem com o intuito de facilitar o seu entendimento. Além disso, o que se pretende, após a apresentação geral, é a criação de uma instância do modelo que ajude na compreensão de todas as etapas.

Os trabalhos identificados na revisão sistemática abordam de forma superficial o processo de seleção de ideias realizado por um especialista, que é o subsídio para um minerador de ideias. O modelo proposto objetiva introduzir critérios utilizados por um especialista humano para separar ideias de textos comuns, através das técnicas *Word Embedding* e *Knowledge Graph*. E, em conjunto com classificadores tradicionais *Suport Verctor Machines*, *Decision Trees*, *Random Forest* e *Naive Bayes*, criar uma ordenação (*ranking*) de textos onde aqueles de maior pontuação possam ser considerados uma ideia. Como resultado final, pretende-se, além do *ranking*, informar a contribuição de cada critério para cada ideia considerando o conjunto em análise.

4.1 APRESENTAÇÃO DO MODELO PROPOSTO

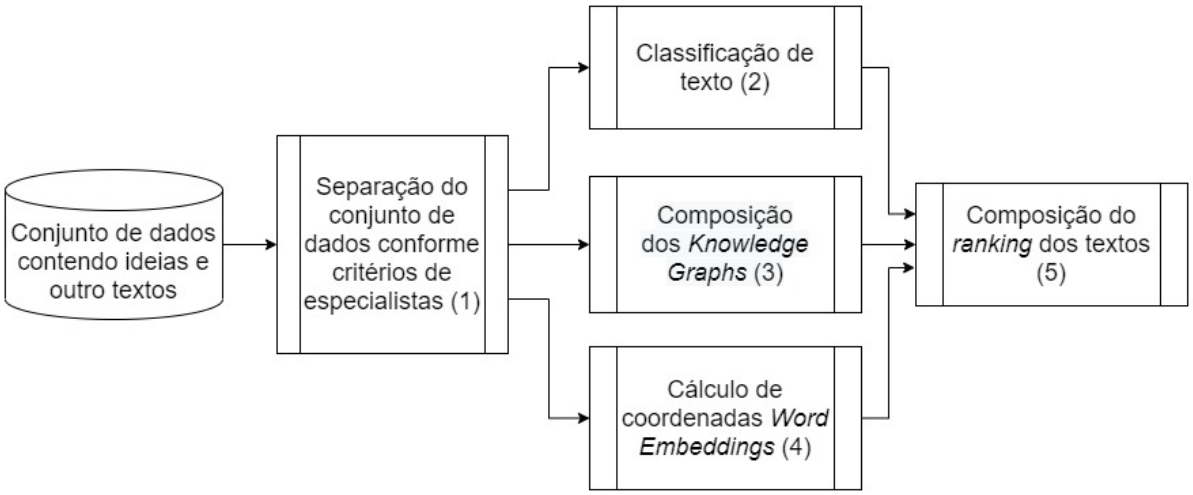
O modelo proposto consiste em uma série de etapas, utilizando os MTECs declarados no referencial teórico, com a finalidade de responder a pergunta de pesquisa e atingir os objetivos geral e específicos da tese. A seguir serão apresentadas as etapas do modelo de forma simplificada e, na seção 4.2, serão explicadas a concepção e funcionalidades de cada etapa de maneira mais detalhada.

A partir do conjunto de dados contendo ideias e textos comuns, as seguintes etapas constituem o modelo:

1) Classificação de texto realizada por um classificador de aprendizado de máquina;
2) Separação do conjunto de dados conforme os critérios dos especialistas;
3) Composição dos *Knowledge Graphs* (KGs), sendo um para cada critério;
4) Cálculo de coordenadas utilizando *Word Embeddings* (WE);
5) Composição do *ranking* dos textos analisados.

A Figura 11 apresenta as etapas do modelo proposto.

Figura 11 – Composição do modelo proposto



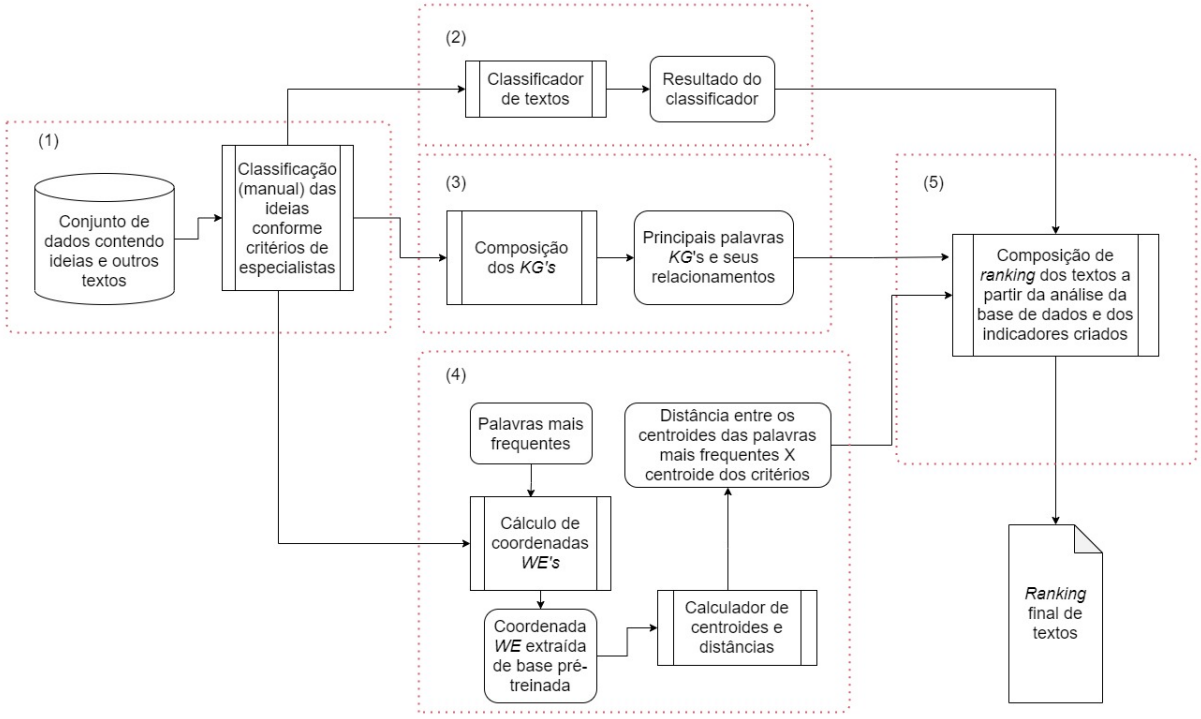
Fonte: autor

O modelo possui, como objetivo final, a composição de um *ranking* de textos levando em conta os resultados de MTECs. Neste sentido, o *ranking* deve apresentar nas posições mais elevadas um conjunto de textos que represente ideias com o intuito de auxiliar especialistas na identificação de quais dessas teriam maior potencial para serem implementadas por determinada organização.

Para tal, são utilizados a classificação de texto, a composição de KGs e o cálculo de coordenadas com base em WEs, através de um treinamento. Este treinamento é realizado por meio de um conjunto de ideias separadas por alguns critérios utilizados por especialistas em suas avaliações. Tal abordagem objetiva simular o comportamento de um especialista no momento em que este classifica determinada ideia, como uma simplificação do processo. O modelo por sua vez incorpora este conhecimento do especialista através do seu treinamento, que por sua vez, adquire características adicionais que o fazem aprimorar o resultado dos classificadores de texto tradicionais.

Considerando as etapas acima descritas, o diagrama geral do modelo proposto é apresentado na Figura 12.

Figura 12 - Diagrama geral do modelo proposto



Fonte: autor

Na seção 4.2 a seguir são detalhadas as etapas do modelo e na seção 4.3 será provida uma exemplificação do funcionamento visando facilitar o entendimento da proposta.

4.2 COMPOSIÇÃO DO MODELO

O modelo, preliminarmente apresentado por meio de uma representação simplificada de sua composição e, na sequência utilizando uma representação mais completa, terá a suas etapas detalhadas nesta seção.

4.2.1 Etapa de separação do conjunto de dados conforme critérios de especialistas

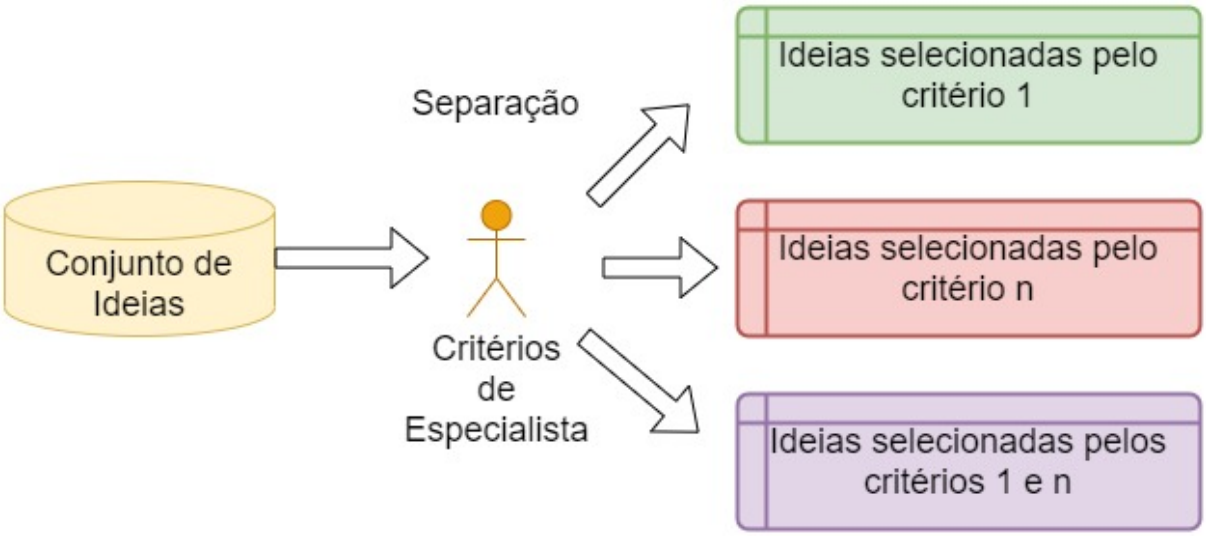
A partir de uma revisão da literatura onde foram selecionados os principais critérios de especialistas utilizados na seleção de ideias, os textos foram lidos e manualmente classificados. Os critérios dos especialistas utilizados foram: originalidade, produtibilidade e viabilidade econômica. Conforme já mencionado na seção 2.2.3, estes critérios fazem parte do aspecto tecnológico, e representam os principais requisitos de escolha neste domínio.

Cabe ressaltar que uma ideia pode ser classificada por um ou mais critérios e isto se reflete no conjunto de treinamento, sendo que ideias que possuem mais de um critério são replicadas nos seus respectivos conjuntos de treinamento. Exemplificando, se um texto contendo uma ideia apresenta os critérios de originalidade e viabilidade econômica, ambos os conjuntos de treinamento de originalidade e viabilidade econômica vão apresentar esta mesma ideia.

Esta classificação tem como principal objetivo a separação de dados para o treinamento, visto que a tarefa de mineração de ideias é um processo complexo se for considerado como um todo. Sendo assim, uma classificação considerando critérios permite tornar o modelo similar ao que realmente é realizado na prática, quando um especialista está avaliando uma ideia.

A seguir é apresentada a Figura 13 que contém o detalhamento da etapa de separação do conjunto de dados conforme critérios de especialistas.

Figura 13 - Detalhamento da etapa de separação do conjunto de dados conforme critérios de especialistas



Fonte: autor

4.2.2 Etapa de classificação de texto

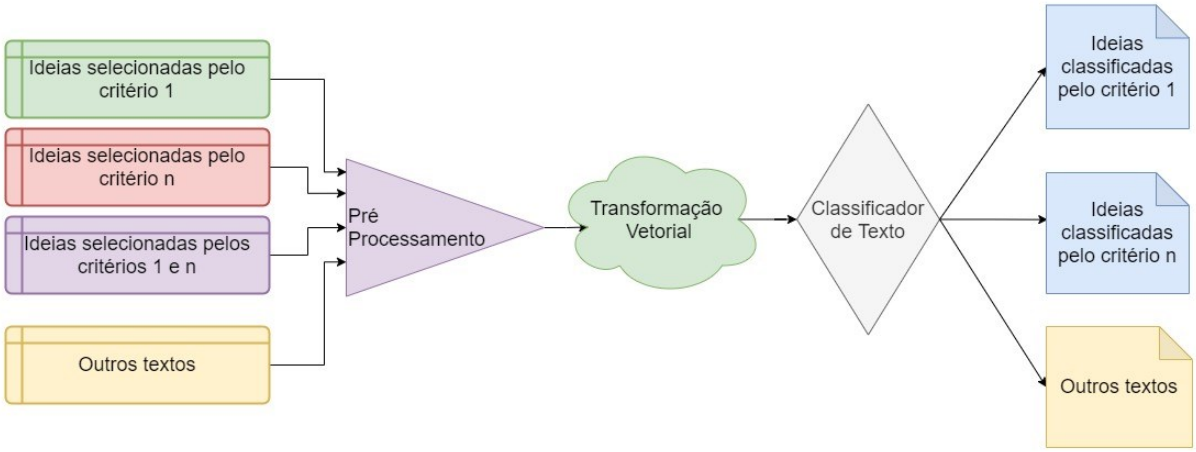
O conjunto de ideias utilizado pelo modelo deve ser constituído de textos contendo ideias selecionadas com potencial de serem implementadas, ou seja, que podem ser desenvolvidas por determinada organização, assim como de textos comuns. Os textos comuns devem ser constituídos de forma aleatória, sem escolher um assunto em particular. Sendo

assim, os classificadores de texto são aplicados e seus resultados participam da composição do *ranking*. De modo geral, qualquer classificador de texto pode ser utilizado nesta tarefa, sendo que aqueles mencionados na seção 2.4.1 serão utilizados por serem amplamente utilizados e citados na literatura aderente a esta tese.

O pré-processamento e a transformação vetorial realizadas na etapa anterior à classificação de texto foram mencionadas nas seções 3.4.3.1 e 3.4.3.2. Como resultado desta etapa, tem-se a separação do conjunto de dados em ideias classificadas por critérios e texto comum. Para tal, o classificador é treinado a partir dos critérios, sendo necessário *n* classificadores treinados (do mesmo tipo) para cada *n* critérios a serem classificados. Exemplificando, na tarefa de classificação de texto pode ser utilizado 3 classificadores NB (NB_O, NB_P e NB_V), sendo o NB_O utilizado para identificar o critério originalidade, o NB_P para identificar o critério produtibilidade e o NB_V para identificar o critério viabilidade econômica.

A seguir a Figura 14 contém o detalhamento da etapa de classificação de texto por classificadores tradicionais de aprendizado de máquina.

Figura 14 – Detalhamento da etapa de classificação de texto



Fonte: autor

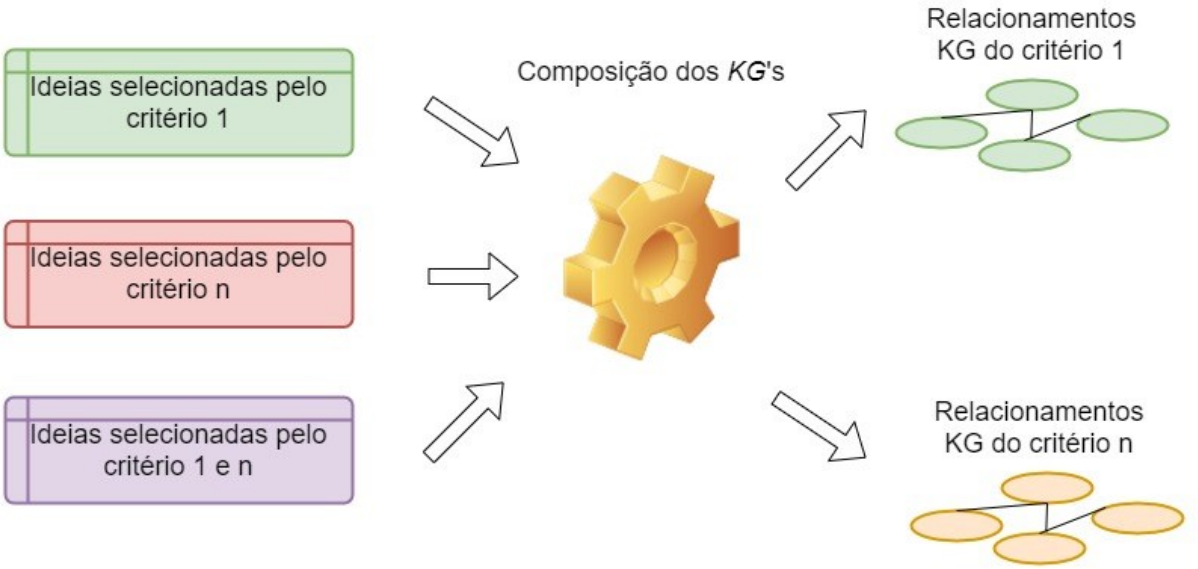
4.2.3 Etapa de composição dos *Knowledge Graphs*

As ideias selecionadas pelos critérios na etapa anterior representam a base de entrada para a composição dos KGs sendo estes responsáveis pela agregação de estruturas de conhecimento ao modelo. As saídas da composição dos KGs farão parte na composição do *ranking* final.

Para o treinamento do modelo, se faz necessária a análise coletiva das ideias classificadas por cada critério. Este conjunto serve de entrada para a composição dos KGs para cada um dos critérios utilizado no modelo. Estes grafos são utilizados como referências ao se analisar um grafo individual de cada texto, ou seja, considerando um determinado grafo de um texto em específico (ideias ou texto comum), este é comparado com o grafo de cada critério de modo que seja possível estabelecer uma aderência no nível de termos e relacionamentos.

A seguir é apresentada a Figura 15 com o detalhamento da etapa de criação de relacionamentos utilizando KGs.

Figura 15 - Detalhamento da etapa composição dos *Knowledge Graphs*



Fonte: autor

4.2.4 Etapa de cálculo de coordenadas utilizando *Word Embeddings*

Neta etapa o conjunto de ideias passa por um processo de identificação das palavras mais relevantes que será a entrada do cálculo de coordenadas WE. Isto é realizado através da retirada de *stop words* e, em seguida, verificando no texto quais são as *n* palavras mais frequentes. Em seguida o cálculo de coordenadas de WEs obtém as coordenadas utilizando-se de uma base pré-treinada no idioma Português, o repositório de WEs do Núcleo Interinstitucional de Linguística Computacional (NILC, 2017).

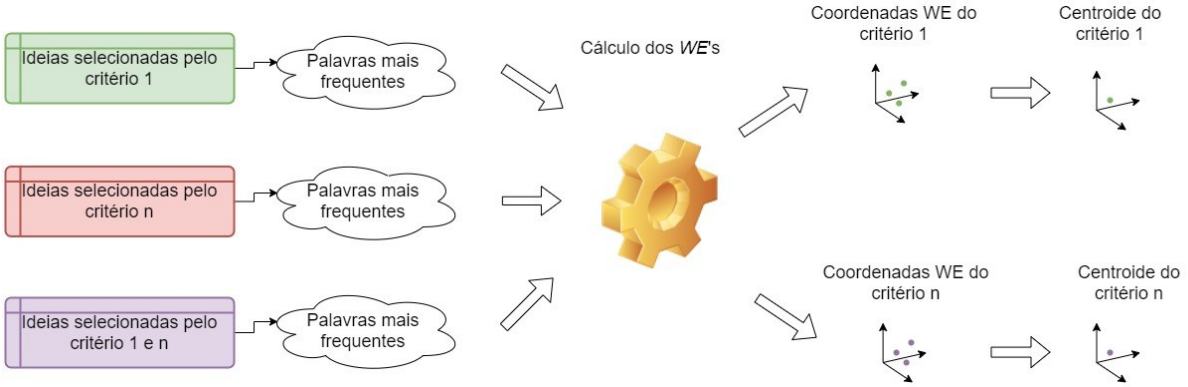
A partir das coordenadas, são realizados os cálculos do centroide geral e dos centroides por critérios. Este centroide geral, considerando todas as ideias serve de referência

para a comparação com o centroide individual de cada texto, permitindo assim, estabelecer a similaridade entre os textos e o centroide de um critério específico. O centroide é obtido geometricamente a partir das coordenadas do WE de cada palavra mais frequente de cada texto. Para fins de cálculos e ilustrações ao longo desta seção, foram consideradas somente 3 dimensões (eixos *x, y* e *z*) para as cinco palavras mais frequentes de ocorrência em determinado texto.

A partir disso, é traçada uma figura geométrica (bloco) de forma que esta consiga englobar os 5 pontos que representam as palavras mais frequentes de ocorrência no texto. O mesmo processo é realizado para todos os textos contendo ideias, sendo então calculado o centroide geral das ideias. Por fim, agrupam-se as ideias selecionadas por cada critério e, a partir disso, calculam-se os centroides das ideias por critério (originalidade, produtibilidade e viabilidade econômica). Para verificar se o texto é uma ideia deve-se calcular a distância entre o centroide geral das ideias e o centroide do texto a ser analisado.

A seguir é apresentada a Figura 16 que contém o detalhamento da etapa de criação de coordenadas utilizando WEs.

Figura 16 - Detalhamento da etapa de cálculo de coordenadas utilizando *Word Embeddings*



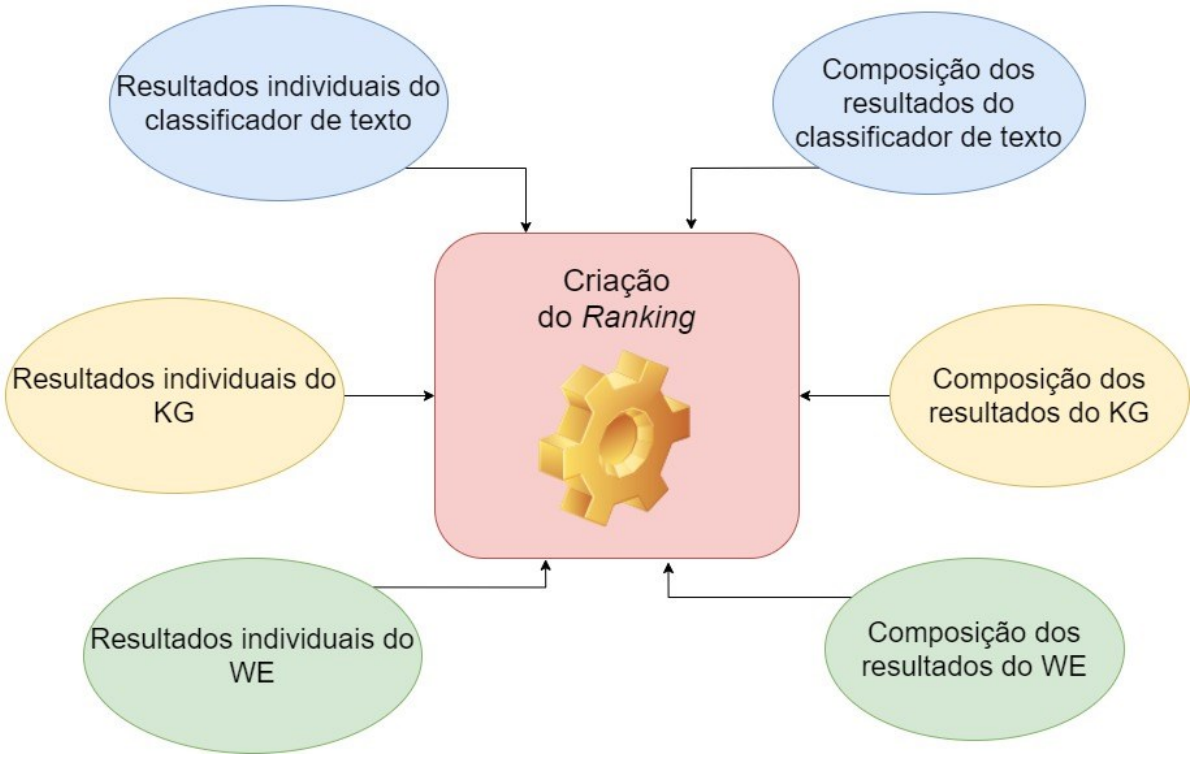
Fonte: autor

4.2.5 Etapa da composição de ordenação (*ranking*) dos textos

Esta etapa é responsável pelo estabelecimento de um *ranking* contendo um grau de pertinência de um texto analisado ser ou não uma ideia, que determina a aderência a cada critério. A composição ocorre pelos resultados descritos nas etapas de classificação de texto, da etapa de composição dos *Knowledge Graphs* e do cálculo de coordenadas utilizando *Word Embeddings*, a partir de seus resultados individuais.

A seguir é apresentada a Figura 17 que detalha a etapa de composição do *ranking* de textos.

Figura 17 - Detalhamento da etapa de composição do *ranking* de textos



Fonte: autor

4.3 EXEMPLIFICAÇÃO DO FLUXO DAS ETAPAS DO MODELO

Para ilustrar o fluxo, será utilizado um texto caracterizado como ideia selecionada a partir do critério **Produtibilidade** para mostrar a sua codificação ao percorrer todas as etapas do modelo. Os demais critérios não serão mostrados, visto que o fluxo seria bastante similar. O conteúdo do texto utilizado no exemplo é apresentado no Quadro7.

Quadro 7 - Texto utilizado para exemplificação do modelo

Criar um APP voltado a gestão de pequenas propriedades rurais com alimentação de dados e controle de desempenho, nas dimensões financeira e produtiva (desempenho produtivo e zootécnico). O objetivo da ideia é facilitar a gestão em propriedades rurais, com ênfase em duas dimensões principais: 1) desempenho econômico: rentabilidade, lucratividade, margem de contribuição, ponto de equilíbrio, indicadores, etc. e 2) desempenho produtivo: identificação e acompanhamento produtivo nas diversas atividades de uma pequena propriedade rural, acompanhando aspectos de produtividade e zootécnicos (controle de animais vacinas peso etc.). A proposta visa oferecer alternativa de gestão por intermédio de um aplicativo que facilite o controle contínuo de dados financeiros e produtivos zootécnicos, considerando custos despesas e receitas da propriedade rural em suas diferentes atividades, gerando informações para a tomada de decisões. Embora haja muita tecnologia e recursos disponíveis no mercado, as pequenas propriedades rurais ainda não dispõem de muitos recursos para o controle de suas atividades, fragilizando a tomada de decisões e em consequência os resultados de suas atividades, uma vez que na grande maioria dos casos não possuem informações minimamente sistematizadas. Dessa perspectiva, resultam também limitações na qualidade de vida das famílias rurais. O desafio pensado segue na linha de uma inovação incremental, onde se pretende superar a falta de alternativas tecnológicas simplificadas de apoio a gestão de pequenas propriedades rurais. O oferecimento da solução de interesse de grande número de pessoas, famílias rurais, as quais em geral já dispõem de smartfones, o que facilitaria o uso de um aplicativo voltado a controlar o desempenho das atividades das propriedades. A contribuição esperada é de importante impacto social e poderá oportuniza maior produtividade e melhor qualidade de vida no campo.

Fonte: Sinapse da Inovação (2018)

Primeiramente, o texto deve ser submetido a um classificador de textos. Para este exemplo utilizou-se o algoritmo *Naive Bayes* (NB), treinado para a identificação de cada critério individualmente. Após a etapa de treinamento o texto foi identificado como uma ideia pelo classificador NB, sendo que este identificou os três critérios no texto.

Na sequência, o texto foi submetido ao codificador WE cujos resultados são apresentados na Tabela 4.

Tabela 4 - Codificação WE do texto exemplo contendo uma ideia

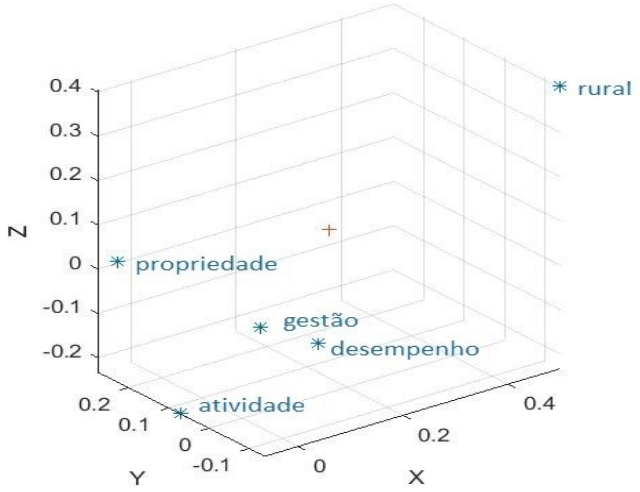
Dimensão	rural	Propriedade	desempenho	atividade	gestão
x	0,500257	-0,027241999	0,068833999	-0,06357	0,196664006
y	-0,13873	0,275985003	-0,09973	0,07132	0,210511997
z	0,406158	0,004379	-0,040605001	-0,23274	-0,193737999

Fonte: autor

A Tabela 4 indica as 5 palavras mais frequentes no texto e suas coordenadas WEs em três dimensões. A partir disso, é traçada uma figura geométrica (bloco) de forma que se

consiga englobar os 5 pontos que representam as palavras. Desta forma, encontra-se o centroide da representação apresentado na Figura 18.

Figura 18 - Coordenadas das palavras mais frequentes do exemplo de ideia e seu centroide



Fonte: autor

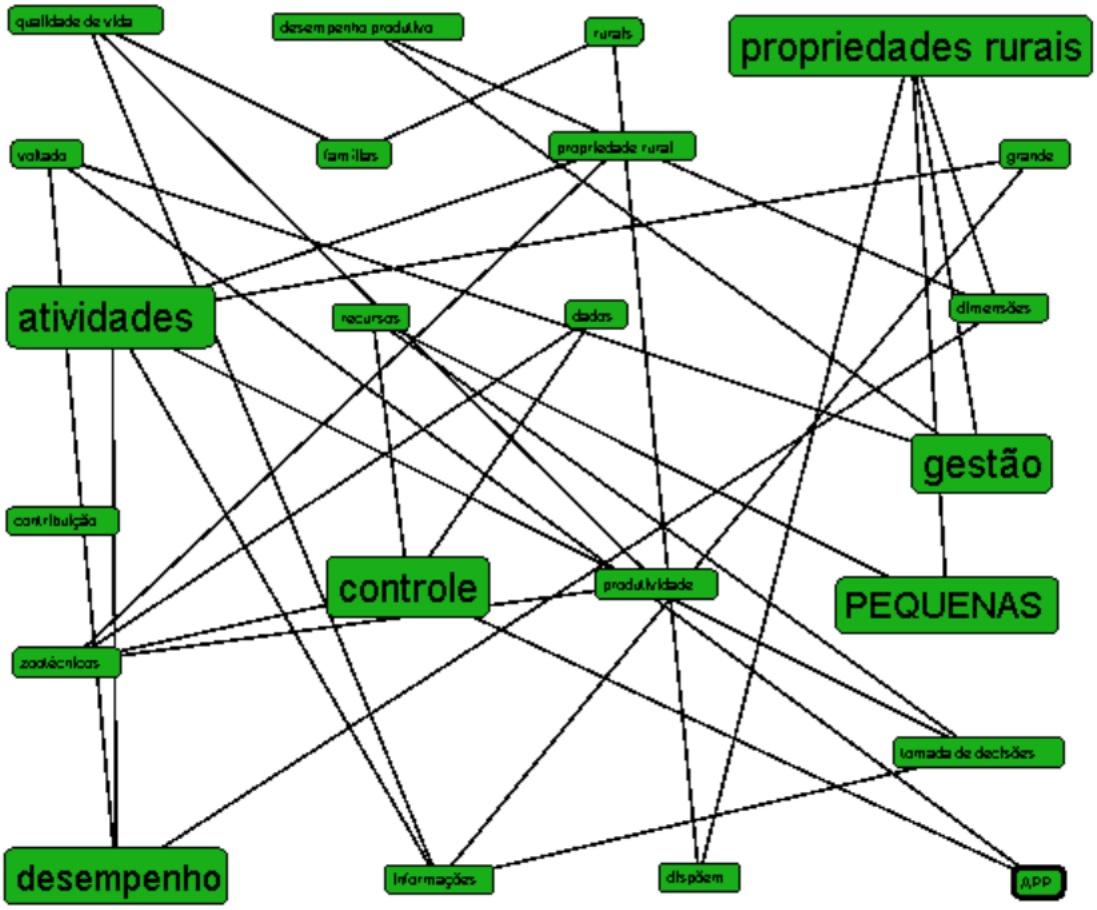
Para o exemplo descrito, o centroide do texto analisado possui as coordenadas (0.2183435, 0.068627502, 0.086709).

Para o cálculo do centroide geral de cada critério é aplicada a Equação 6 (apresentada em 4.2.4) a todos os textos que apresentam originalidade, produtibilidade e viabilidade econômica. As coordenadas do centroide do critério originalidade são (0.212572992, 0.1907565, 0.155280992), do critério produtibilidade são (0.172589481, 0.27659753, 0.123676986) e do critério viabilidade econômica são (0.159202501, 0.283354044, 0.245565489).

No exemplo descrito, a distância entre o centroide do texto analisado e do critério originalidade foi 0.197638502, do critério produtibilidade foi 0.246253852 e do critério viabilidade econômica foi de 0.331523084. No treinamento realizado com todos os textos, são estabelecidos os limites de aderência a cada critério. Considerando os WEs de cada critério, o texto foi classificado com os critérios Originalidade e Produtibilidade.

Em continuidade, é aplicado o método de composição do KG, onde são identificados os principais termos e os seus relacionamentos. Esta etapa produz um KG para cada critério considerado no modelo, bem como um grafo individual para cada texto do conjunto de dados. A Figura 19 apresenta um grafo de exemplo para o critério produtibilidade.

Figura 19 – Grafo de uma ideia selecionada pelo critério produtibilidade



Fonte: autor

Os conceitos que coincidem com o KG do critério produtibilidade foram gestão e desempenho, com a relação mais próxima encontrada sendo, neste caso, entre gestão e desempenho. Verificou-se ainda que foram encontradas correspondências entre as relações do texto com as relações gerais do critério produtibilidade.

A seguir na Tabela 5 é mostrado um fragmento do ranking, ilustrando o cálculo descrito em 4.2.5.

Tabela 5 – Fragmento do *ranking* de textos

índice	NB_O	NB_P	NB_V	KG_O	KG_P	KG_V	WE_O	WE_P	WE_V	Res.NB	Res.KG	Res.WE	cálculo	G(O)%	G(P)%	G(V)%	ranking
7	1	1	1	1	1	1	1	1	0	1	1	1	95,5	37,5	37,5	25,0	20

Fonte: autor

Neste exemplo a posição do *ranking* foi a 20, sendo que o cálculo geral para composição do ranking foi de 95.55, tendo o texto sido classificado entre os 122 primeiros, ou seja, está dentro do conjunto das ideias. Porém, seus graus de pertinência não foram exatamente correspondentes à classificação inicial que era referente ao critério produtibilidade, mas sim 38% de produtibilidade, 38% de originalidade e 25% de viabilidade econômica.

Por fim, o especialista terá informações detalhadas sobre os resultados de cada etapa do modelo, assim como a ordem (*ranking*) sugerida dos textos (ideias ou não) em análise. A partir deste *ranking*, determinado especialista pode concentrar esforços nos textos que obtiveram as melhores classificações com o objetivo de identificar mais claramente quais desses são ideias e se essas possuem potencial de implementação/desenvolvimento.

4.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o modelo proposto constituído com base na problemática descrita no Capítulo 1 e a partir do referencial teórico apresentado no Capítulo 2. Foram detalhadas cada etapa do modelo e a função das mesmas, apresentando figuras que pudessem melhor explicar o seu funcionamento.

Ressalta-se ainda que foram declaradas todas as variáveis utilizadas na produção do cálculo do *ranking*, bem como as suas equações. Para os graus de pertinência a cada critério foram declaradas as variáveis utilizadas e os seus cálculos.

Por fim, como forma de exemplificação da aplicação do modelo, foi utilizado um texto específico do conjunto de dados e apresentado todo o fluxo dos cálculos para se chegar no *ranking* e aos graus de pertinência de cada critério.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste capítulo serão apresentados na seção 5.1 os resultados das avaliações das etapas do modelo no sentido de definir as melhores técnicas para a sua instanciação, utilizando o conjunto de dados definido em 3.4.4.1. Após isto, o modelo será efetivamente avaliado na seção 5.2 com o conjunto de dados definido em 3.4.4.2 sendo os resultados discutidos na sequência.

5.1 ANÁLISE DAS ETAPAS DO MODELO

Nesta seção serão detalhados os testes de avaliação de cada etapa do modelo proposto.

5.1.1 Avaliação da Classificação de Ideias por Critérios utilizando Classificadores

Para a elaboração da avaliação da etapa de classificação de texto pelos classificadores tradicionais NB, SVM, DT e RF, foram definidas as seguintes etapas:

1) Utilização do conjunto de dados definido em 3.4.4.1;
2) Submissão do conjunto de dados aos classificadores de texto NB, SVM, DT e RF para testar a capacidade dos classificadores em separar conteúdos que representem ideias de textos comuns. Conforme indicado na seção 3.4.3.3, utilizou-se aqui o procedimento de validação cruzada;
3) Nova submissão do conjunto de dados aos classificadores NB, SVM, DT e RF para testar a capacidade dos classificadores em separar cada critério individual dos especialistas definidos no item 2;
4) Comparação dos resultados obtidos (por meio das métricas acurácia, precisão, revocação e coeficiente *kappa*) sobre o conjunto de dados de ideias considerando os itens 2 e 3 com a finalidade de escolher a melhor abordagem para a classificação das ideias e definir dois classificadores com os melhores resultados para a composição do modelo final.

Para a realização dos testes, foram utilizados os critérios definidos na seção 2.2.3, sendo a originalidade, a produtibilidade e a viabilidade econômica.

A primeira execução dos classificadores NB, SVM, DT e RF levou em consideração o conjunto de dados elaborado para este estudo. Tem por finalidade a verificação do desempenho dos classificadores na separação de ideias e de textos comuns. A seguir são apresentadas as matrizes confusão, conforme Tabela 6, gerada a partir desta execução.

Tabela 6 - Matrizes confusão considerando a separação em ideias e textos comuns

(a)

Matriz confusão: Naive Bayes		
Atual / Predito	Ideias	Texto Comum
Ideias	97	25
Texto Comum	21	79

(b)

Matriz confusão: Support Vector Machine		
Atual / Predito	Ideias	Texto Comum
Ideais	115	7
Texto Comum	19	81

(c)

Matriz confusão: Decision Trees		
Atual / Predito	Ideias	Texto Comum
Ideias	97	25
Texto Comum	21	79

(d)

Matriz confusão: Random Forest		
Atual / Predito	Ideias	Texto Comum
Ideias	121	1
Texto Comum	31	69

Fonte: autor

As Tabela 6a até 6d, mostram que os classificadores foram capazes de separar bem os textos comuns dos textos que representam ideias. Cada matriz confusão destaca em azul os acertos realizados pelos classificadores.

A segunda execução dos classificadores NB, SVM, DT e RF também levou em consideração o conjunto de dados produzido para este estudo. Porém, analisando a capacidade dos classificadores em separar as ideias conforme os três critérios individualmente estabelecidos comparando-se aos textos comuns.

A finalidade deste teste foi verificar o desempenho dos classificadores na tarefa de separação em duas classes distintas, ou seja, cada critério individualmente comparado aos

textos comuns. Para composição do resultado são considerados os acertos dos classificadores individuais treinados para identificar cada critério, onde considera-se acerto se os resultados obtidos alcancem a condição de acerto de pelo menos um dos critérios. Para exemplificar, considerando um texto definido como ideia que possua 1 ou mais critérios identificados na etapa 4.2.1, a composição do resultado considera acerto caso pelo menos 1 critério seja identificado pelos classificadores treinados individualmente para cada critério.

Segue abaixo as matrizes confusão, conforme Tabela 7a até 7d, geradas a partir desta segunda execução.

Tabela 7 - Matrizes de confusão considerando as ideias separadas por critério versus texto comum

(a)

Matriz confusão: Naive Bayes		
Atual / Predito	Ideias separadas	Texto Comum
Ideias separadas	120	2
Texto Comum	0	100

(b)

Matriz confusão: Support Vector Machine		
Atual / Predito	Ideias separadas	Texto Comum
Ideias separadas	113	9
Texto Comum	2	98

(c)

Matriz confusão: Decision Trees		
Atual / Predito	Ideias separadas	Texto Comum
Ideias separadas	111	11
Texto Comum	17	83

(d)

Matriz confusão: Random Forest		
Atual \ Predito	Ideias separadas	Texto Comum
Ideias separadas	118	4
Texto Comum	42	58

Fonte: Autor

Ao comparar as matrizes de confusão da Tabela 6 e da Tabela 7, verificou-se que há uma melhora de desempenho na segunda execução. Isto mostra que os classificadores analisados foram capazes de separar as ideias dos textos comuns e conseguiram melhorar o índice de acerto quando treinados para encontrar critérios.

Existem na literatura alguns trabalhos, dentre os quais podem ser citados, Badawi e Altinçay (2014), Elhassan e Ali (2019) e Li, Liu e Ng (2010), em que os classificadores treinados para escolhas binárias apresentam melhor desempenho comparados com a utilização de mais de duas classes de escolha. A complexidade de uma classificação é maior conforme o aumenta o número de classes (ELHASSAN; ALI, 2019).

Finalmente, é apresentada a Tabela 8 contendo os índices de acurácia, precisão, revocação, F1-*Score* e o coeficiente *kappa*, onde os desempenhos de todas as execuções podem ser analisados e comparados entre si.

Tabela 8 - Resultados da acurácia, precisão, revocação, F1-*Score* e coeficiente *kappa* para os classificadores NB, SVM, DT e RF nas duas execuções

Classificadores	Métricas avaliadas	Ideais (1ª execução)	Ideias com os critérios (2ª execução)
NB	*Acurácia*	0,793	0,991
	Precision	0,822	1,000
	Recall	0,795	0,984
	F1-Score	0,808	0,992
	Kappa	0,583	0,982
SVM	*Acurácia*	0,883	0,950
	Precision	0,858	0,983
	Recall	0,943	0,926
	F1-Score	0,898	0,954
	Kappa	0,761	0,901
DT	*Acurácia*	0,793	0,874
	Precision	0,822	0,867
	Recall	0,795	0,910
	F1-Score	0,808	0,888
	Kappa	0,583	0,744
RF	*Acurácia*	0,856	0,793
	Precision	0,796	0,738
	Recall	0,992	0,967
	F1-Score	0,883	0,837
	Kappa	0,701	0,567

Fonte: Autor

Os resultados demonstram que os classificadores *Naive Bayes*, *Support Vector Machines*, *Decision Trees* e *Random Forest* apresentam desempenho similar considerando a tarefa de separar os textos contendo ideias dos textos comuns (1ª execução), tendo o SVM atingido os melhores resultados. Quando os critérios foram considerados individualmente (2ª

execução), os classificadores apresentam desempenho superior se comparado às execuções anteriores. Cabe ressaltar que, quando os critérios são separados, os classificadores NB e SVM tiveram desempenho superior aos classificadores DT e RF.

A partir desta análise é possível verificar que a melhor abordagem consiste na realização da classificação de textos utilizando cada critério de um especialista de forma individual, sendo que os classificadores NB e SVM obtiveram os melhores desempenhos. Nestes casos, o coeficiente *kappa* encontrado variou entre 0.901 a 0.982, indicando um desempenho excelente de acordo com Landis e Koch (1977).

Cabe ressaltar que o conjunto de dados utilizado no treinamento dos classificadores é pequeno e homogêneo, o que pode ser um dos motivos para os elevados índices atingidos. Sabe-se que em condições reais, os dados podem ser esparsos e incompletos, gerando assim uma queda neste desempenho. Porém, a constatação de que os classificadores NB e SVM são capazes de separar adequadamente textos que representam ideias de textos comuns fica comprovada pelos testes realizados possibilitando a utilização do classificador NB como referência na composição do modelo proposto na tese.

5.1.2 Avaliação da Separação de Textos utilizando *Word Embeddings*

Para o desenvolvimento dos testes de avaliação a respeito da separação de textos utilizando *Word Embeddings* (WE*)* foram definidas as seguintes etapas:

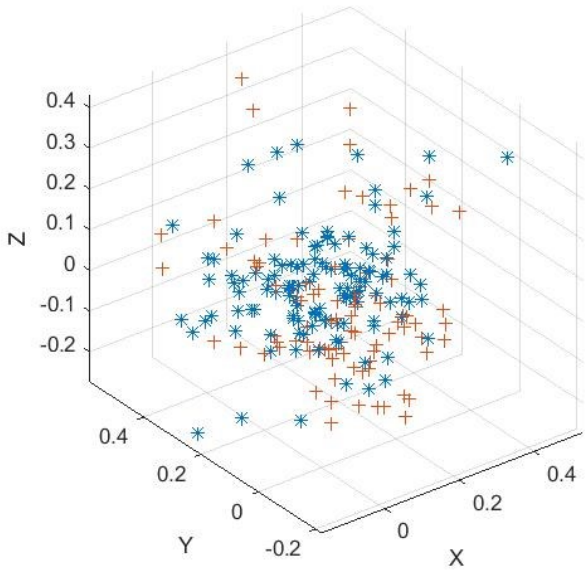
1) Utilização do conjunto de dados definido em 3.4.4.1;
2) Realizar o pré-processamento retirando caracteres especiais e *stopwords* do conjunto de dados;
3) Extração das 5 palavras de maior frequência de ocorrência em cada texto do conjunto de dados. Através de testes foi verificado que a melhor configuração a ser utilizada foi de 5 palavras de maior frequência, sendo que abaixo deste número o desemprenho diminui e acima, o desempenho não sofria alterações até o número 8 onde, a partir deste ponto, o desemprenho também diminui.
4) Submissão das palavras a um vocabulário previamente treinado através de WE, tendo como resultado a sua codificação equivalente. Conforme indicado na seção 3.4.3.3, utilizou-se aqui o procedimento de validação cruzada;

5) Simplificação do grau da codificação das palavras obtidas para k=3 com a finalidade de visualização posterior em 3 dimensões. Desta forma, cada texto do conjunto de dados foi resumido em n=5 pontos de k=3 dimensões;
6) Encontrar uma figura geométrica (bloco) mínima, formada pelos pontos de cada texto e seu centroide;
7) Plotagem dos centroides calculados, de forma que se possa separar por cores os textos contendo ideias e os textos comuns;
8) Plotagem dos centroides de cada critério para demonstração de separabilidade e estabelecimento através de testes dos limites de separabilidade de cada critério;
9) Cálculo de desempenho considerando a separação entre cada critério versus textos comuns.

Para implementação dos WEs foi utilizado o *Word2Vec*, que é um método de incorporação de palavras proposto por Mikolov *et al.* (2013). Tem como princípio a aprendizagem de vetores dimensionais usando um dos dois modelos neurais distintos: *Continuous Bag of Words* (CBOW) ou *Skip-Gram*. Neste trabalho, optou-se pela utilização do CBOW, que de acordo com Mikolov *et al.* (2013) é mais rápido e funciona bem com palavras frequentes.

Primeiramente, são calculadas as codificações dos WEs para as 5 palavras mais frequentes de cada texto. A partir disso, é traçada então uma figura geométrica (bloco) de forma que se consiga englobar minimamente n pontos (palavras). Encontra-se então o centroide desta figura. Este procedimento é realizado para cada um dos 122 textos contendo ideias e também os 100 textos comuns. O resultado é apresentado na Figura 20, sendo os textos comuns representados pelo símbolo '+' em vermelho e os textos com ideias representados pelo símbolo '*' em azul.

Figura 20 - Plotagem das ideias e textos comuns



Fonte: autor

A partir deste gráfico pode-se verificar que há uma maior proximidade entre os pontos correspondentes aos textos que contêm ideias, demonstrando ser possível uma diferenciação matemática dos textos. Para comprovar tal verificação visual foi encontrado o ponto médio dos centroides das ideias e calculadas as médias das distâncias entre o centroide de cada ideia versus seu ponto médio, onde o resultado das médias foi de 0.092. O mesmo foi efetuado para os textos comuns e seu resultado foi 0.109. Esta diferenciação entre as médias mostra que os centroides dos textos contendo ideias apresentam uma menor distância entre si, mostrando-se mais densos.

5.1.3 Avaliação da Criação de *Knowledge Graphs*

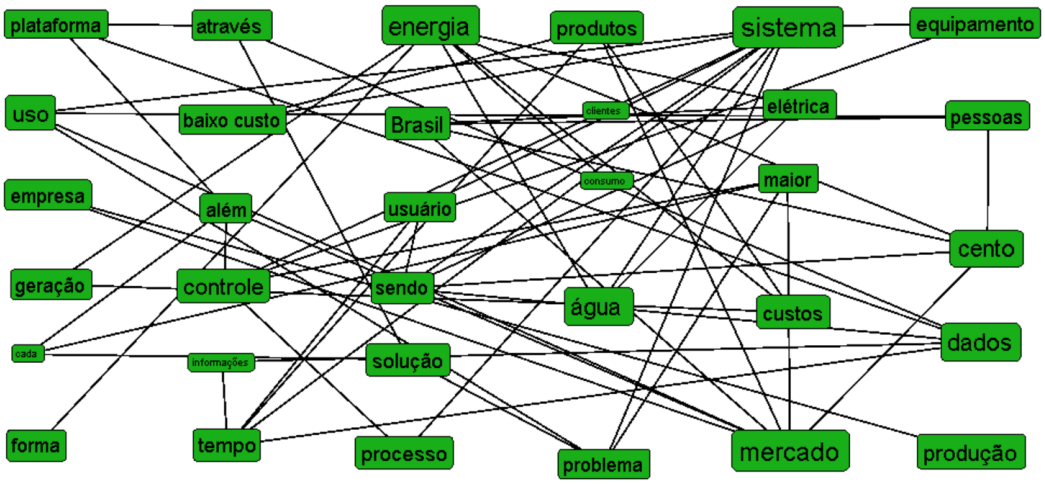
Para o desenvolvimento dos testes de avaliação da etapa baseadas no conceito de *Knowledge Graph* (KG) foram definidos os seguintes passos:

1) Utilização do conjunto de dados definido em 3.4.4.1;
2) Submissão do conjunto de dados separado por critérios ao método de geração de KGs, onde é gerado um grafo geral de cada critério;
3) Verificação da aderência considerando os termos e os relacionamentos em comuns entre e os grafos individuais e o grafo de cada critério.

O conjunto de dados separados por critérios declarado na seção 3.4.4.1 foi submetido à aplicação *Sobek Mining*® que criou um grafo (neste trabalho referenciado como KG) para cada critério. Os KGs são capazes de exibir as relações existentes nos textos de cada critério, podendo então ser considerados como um treinamento para uso posterior. O objetivo nesta etapa consiste em comparar um texto específico (ideia ou não), representado na forma de um grafo, ao KG de cada critério, verificando os termos e relacionamentos em comum. Quanto maior a quantidade de termos (ou também chamados de conceitos pelo *Sobek Mining*®) e relacionamentos em comum, maior a possibilidade deste texto ser atribuído para um determinado critério.

O KG gerado para o critério Viabilidade Econômica limitado a 50 conceitos, conforme declarado em 3.4.3, é mostrado a seguir na Figura 21.

Figura 21 - KG do critério Viabilidade Econômica



Fonte: autor

Os vinte termos (conceitos) mais frequentes são mostrados no Quadro 8.

Quadro 8 - Conceitos mais frequentes para o critério Viabilidade Econômica

Conceito selecionado	Número de vezes que o conceito aparece nos textos
sistema	66
mercado	60
energia	51
cento	46
dados	42
água	42
produção	39
uso	39
controle	36
Brasil	33
produto	30
solução	30
tempo	30
equipamento	29
custo	28
processo	28
baixo custo	25
geração	25
maior	25
problema	23

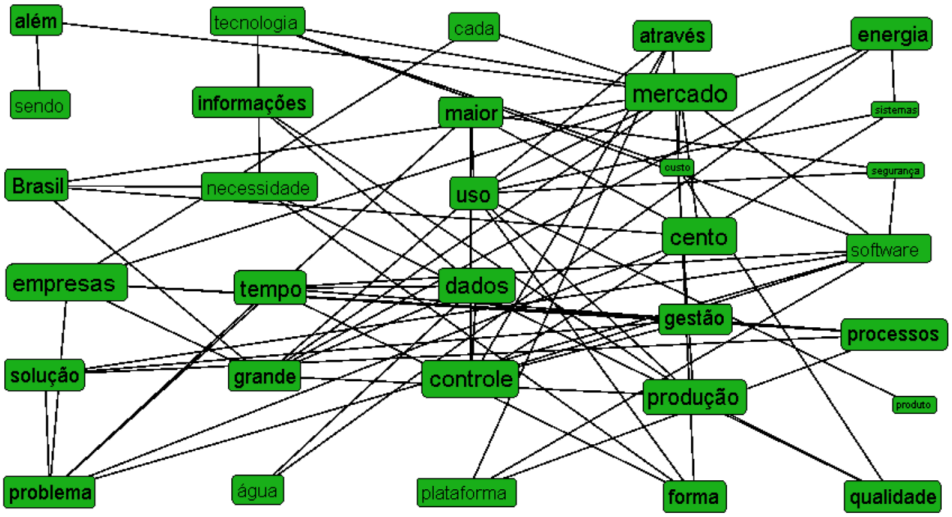
Fonte: autor

As relações mais relevantes, retiradas a partir da análise manual do grafo KG, que remetem ao critério viabilidade identificadas a partir do KG são:

- baixo custo – produção - solução
- baixo custo – sistema - desenvolvimento
- controle - solução
- geração - tempo
- problema - solução
- desenvolvimento - produto

O KG gerado para o critério Produtibilidade é mostrado a seguir na Figura 22.

Figura 22 - KG do critério Produtibilidade



Fonte: autor

Os vinte termos (conceitos) mais frequentes são mostrados no Quadro 9.

Quadro 9 - Conceitos mais frequentes para o critério Produtibilidade

Conceito selecionado	Número de vezes que o conceito aparece nos textos
mercado	45
empresa	40
cento	39
controle	38
dados	37
uso	33
tempo	31
produção	30
energia	27
Brasil	26
maior	25
forma	24
processo	24
solução	24
informação	23
gestão	22
grande	22
qualidade	21
problema	21
necessidade	18

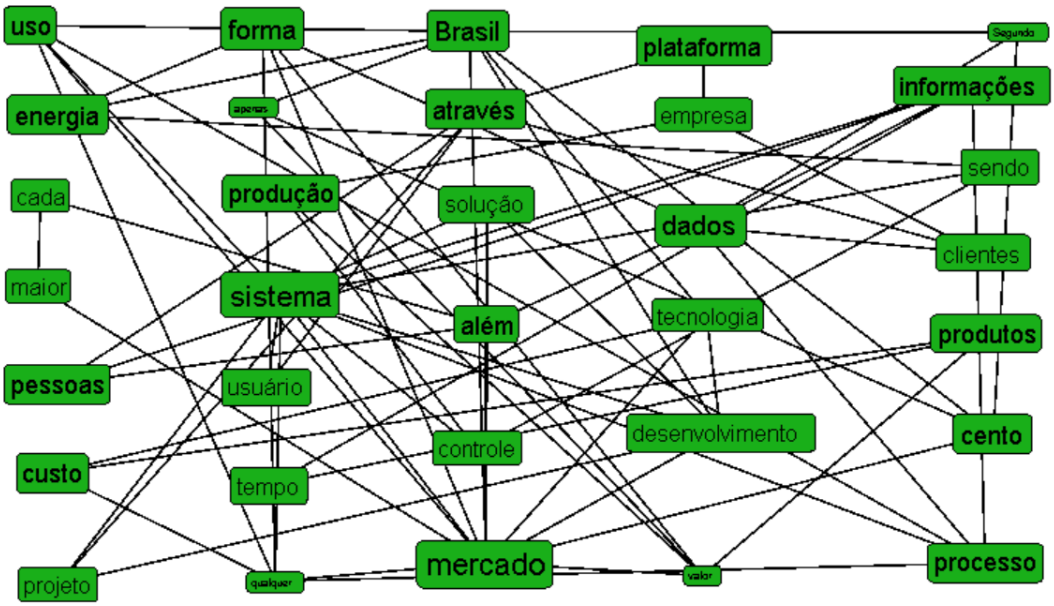
Fonte: autor

Já as relações mais relevantes que remetem ao critério de produtibilidade a partir do KG são:

- aumento – produção - qualidade
- grande - produção
- produção – qualidade - processo
- maior - tempo
- necessidade - qualidade
- solução - processo
- maior - controle

O KG gerado para o critério Originalidade é mostrado a seguir na Figura 23.

Figura 23 - KG do critério Originalidade



Fonte: autor

Os vinte termos (conceitos) mais frequentes são mostrados no Quadro 10.

Quadro 10- Conceitos mais frequentes para o critério Originalidade

Conceito selecionado	Número de vezes que o conceito aparece nos textos
dado	35
forma	35
mercado	32
plataforma	32
através	30
informação	30
processo	27
Brasil	26
pessoa	23
solução	21
cento	20
uso	19
energia	18
tempo	18
clientes	17
desenvolvimento	17
produto	17
ideia	16
produção	16
inovação	15

Fonte: autor

Por fim, as relações mais relevantes que remetem ao critério de originalidade a partir do KG são:

- desenvolvimento - solução
- desenvolvimento - sistema
- projeto – sistema - dado
- problemas - ideias
- processo - ideia
- produção - produto
- dado – informação – gestão

O KG é um método que consegue eliminar 100% dos textos comuns para o conjunto de dados de testes, declarado em 3.4.4.2, pois os mesmos não apresentaram qualquer relação em comum com os textos que indiquem ideias. Outra grande vantagem é a sua facilidade de promover um treinamento contínuo, pois os resultados, após validados por um especialista,

podem servir de entrada para incrementar o KG de cada critério. Por estas condições, os KGs foram utilizados na proposição do modelo para a composição do *ranking*.

5.1.4 Avaliação do *Ranking*

O *ranking* é criado realizando os cálculos descritos em 4.2.5. Para o cálculo do *ranking* foram obtidas a média aritmética entre a composição dos resultados dos classificadores, dos WEs e dos KGs e dos resultados individuais dos classificadores, dos WEs e dos KGs, conforme equação 2.

Pelo fato deste não ser o *ranking* final, são apresentados apenas os acertos obtidos a partir de determinada quantidade de dados, indicados na Tabela 9.

Tabela 9 – Avaliação de desempenho do *ranking*

Posição	Acertos	Porcentagem de acertos
99	99	100%
122	114	93%

Fonte: autor

Os dados da tabela mostram que as primeiras 99 posições do *ranking* apresentaram acerto total. Na posição 122 que representa todo o conjunto de ideias, o acerto foi de 93%.

5.2 VERIFICAÇÃO DO MODELO

Nesta seção serão detalhados os testes finais do modelo proposto, considerando os resultados obtidos nos testes descritos na seção 5.1, utilizando o conjunto de dados descrito em 3.4.4.2.

Conforme mencionado na seção anterior, utilizou-se inicialmente o conjunto de dados definido em 3.4.4.1 como conjunto de treinamento para todas os MTECs considerados na avaliação do modelo. Ou seja, este conjunto de dados serviu de base para produção dos modelos dos classificadores (DT, NB, SVM e RF) para a uma avaliação inicial e para a geração definitiva dos WEs e KGs.

Durante esta fase de teste verificou-se que os WEs e KGs, produzidos a partir do primeiro conjunto de dados, conseguiram capturar adequadamente os critérios utilizados por especialistas (originalidade, produtibilidade e viabilidade econômica) e, portanto, foram mantidos como elementos de referência para a avaliação final do modelo.

Por outro lado, ao considerar todo o segundo conjunto de dados (definido em 3.4.4.2) como conjunto de verificação, os resultados não se mostraram adequados, atingindo uma acurácia muito baixa. A justificativa pode ser encontrada em trabalhos como os de Chen e Cardie (2018) e Zou, Yang e Wu (2021), ao afirmarem que classificadores de texto tendem a ser dependentes de domínio e, em muitos casos, terão o seu desempenho degradado em domínios que sejam diversos àqueles utilizados durante o treinamento

De modo a evitar o problema acima mencionado e, ainda, minimizar a ocorrência de *overfitting* dos classificadores, utilizou-se, para o segundo conjunto de dados (seção 3.4.4.2) a separação de 80% deste conjunto (454 elementos) para treinamento do classificador e 20% deste conjunto (122 elementos) para os testes do classificador em si e, por consequência, a elaboração final do *ranking*. Esta divisão foi realizada de forma aleatória e balanceada, mantendo a proporção entre ideias e projetos. Vale salientar que a etapa de treinamento seguiu as definições apresentadas na seção 3.4.3.3.

Este conjunto de dados, denominado aqui de conjunto de verificação do modelo , foi submetido ao classificador NB com a configuração de identificar cada critério. O mesmo conjunto também foi submetido aos KGs e WEs com a configuração de identificação individual de critérios. Cabe ressaltar que foi também realizado um treinamento dos KGs e WEs a partir deste conjuntos de verificação do modelo e os resultados obtidos foram muito próximos ao treinamento utilizado o conjunto de dados descrito em 3.4.4.1. Devido a esta capacidade de generalização dos KGs e WEs, manteve-se o treinamento indicado descrito em 5.1.2 e 5.1.3.

A partir disso, foi possível gerar o *ranking* das ideias, onde são calculadas a posição e o grau de pertinência aos critérios, para cada texto. Além de avaliar a capacidade de separar uma ideia de um texto que não deva ser considerado uma ideia, o *ranking* é capaz de fornecer um grau de pertinência a cada critério considerando os resultados individuais dos classificadores, dos WEs e dos KGs. Este *ranking* está apresentado de duas formas nos apêndices. No Apêndice A são apresentadas todas as variáveis utilizadas para a realização dos cálculos, bem como seus resultados finais e os graus de pertinência a cada critério. O *ranking* é apresentado na ordem dos índices dos textos. Para facilitar a visualização, os resultados são então ordenados de forma decrescente, para poder ter acesso aos valores do topo do *ranking* nas primeiras posições da tabela. Isto está apresentado no Apêndice B. A Tabela 10 a seguir apresenta um fragmento do *ranking* com as 50 primeiras posições.

Tabela 10 - *Ranking* final considerando os primeiros 50 textos

Posição no ranking	Valor do ranking	Índice	Ideia (1) / Projeto (0)	Grau de pertinência (O)	Grau de pertinência (P)	Grau de pertinência (V)
1	100,00	5	1	33,33%	33,33%	33,33%
2	100,00	13	1	33,33%	33,33%	33,33%
3	100,00	31	1	33,33%	33,33%	33,33%
4	100,00	34	1	33,33%	33,33%	33,33%
5	100,00	42	1	33,33%	33,33%	33,33%
6	100,00	44	1	33,33%	33,33%	33,33%
7	100,00	52	1	33,33%	33,33%	33,33%
8	100,00	55	1	33,33%	33,33%	33,33%
9	91,67	12	1	33,33%	22,22%	33,33%
10	91,67	20	1	33,33%	22,22%	33,33%
11	91,67	26	1	22,22%	33,33%	33,33%
12	91,67	32	1	33,33%	33,33%	22,22%
13	91,67	45	1	33,33%	33,33%	22,22%
14	83,33	8	1	22,22%	33,33%	22,22%
15	83,33	9	1	22,22%	33,33%	22,22%
16	83,33	28	1	22,22%	33,33%	22,22%
17	83,33	41	1	22,22%	33,33%	22,22%
18	83,33	43	1	22,22%	33,33%	22,22%
19	83,33	46	1	22,22%	33,33%	22,22%
20	83,33	50	1	22,22%	33,33%	22,22%
21	75,00	27	1	22,22%	22,22%	22,22%
22	75,00	47	1	22,22%	22,22%	22,22%
23	75,00	51	1	22,22%	22,22%	22,22%
24	75,00	54	1	33,33%	11,11%	22,22%
25	66,67	2	1	11,11%	22,22%	22,22%
26	66,67	4	1	22,22%	22,22%	22,22%
27	66,67	7	1	22,22%	22,22%	22,22%
28	66,67	10	1	22,22%	22,22%	22,22%
29	66,67	14	1	22,22%	22,22%	22,22%
30	66,67	15	1	22,22%	22,22%	22,22%
31	66,67	19	1	22,22%	22,22%	22,22%
32	66,67	21	1	22,22%	22,22%	22,22%
33	66,67	22	1	22,22%	22,22%	22,22%
34	66,67	23	1	22,22%	22,22%	11,11%
35	66,67	33	1	22,22%	22,22%	22,22%
36	66,67	37	1	22,22%	22,22%	22,22%
37	66,67	39	1	22,22%	22,22%	22,22%
38	66,67	48	1	22,22%	22,22%	22,22%
39	66,67	64	0	22,22%	22,22%	22,22%
40	66,67	101	0	22,22%	22,22%	22,22%
41	66,67	110	0	22,22%	22,22%	22,22%

42	58,33	3	1	22,22%	11,11%	22,22%
43	58,33	6	1	22,22%	11,11%	22,22%
44	58,33	11	1	11,11%	22,22%	22,22%
45	58,33	16	1	22,22%	22,22%	11,11%
46	58,33	17	1	22,22%	11,11%	22,22%
47	58,33	24	1	22,22%	22,22%	11,11%
48	58,33	30	1	22,22%	11,11%	22,22%
49	58,33	35	1	22,22%	11,11%	22,22%
50	58,33	36	1	22,22%	11,11%	22,22%

Fonte: autor

Analisando os resultados do *ranking* foi elaborada uma análise com o objetivo de clarificar a acurácia em diferentes pontos de corte. Para tal, foi definido a cada 10 posições do *ranking* geral (Apêndice B) o nível de acurácia, ou seja, a quantidade de acertos considerando a quantidade de textos até o ponto determinado. A Tabela 11 apresenta os resultados.

Tabela 11 - Avaliação de desempenho do *ranking* final a cada 10 textos

Posição	Acertos	Porcentagem de acertos
10	10	100%
20	20	100%
30	30	100%
40	38	95%
50	47	94%
56	50	89%
60	50	83%
70	53	76%
80	54	68%
90	55	61%
100	56	56%
112	56	50%

Fonte: autor

A partir desta tabela é possível observar que as primeiras 30 posições o *ranking* possui um acerto de 100% Este valor decresce para 95% se consideradas as 40 primeiras posições, sendo que na posição 56 (número de ideias presentes no conjunto de testes referente ao conjunto definido em 3.4.4.2), onde uma ordenação ideal atingiria 100%, o acerto foi de 89%. A partir da posição 56 são obtidos os erros que finalizam com 50% na posição 112 (Tabela 11), condição esta já esperada visto que o conjunto de testes possui o mesmo número de ideias e projetos gerais.

Similarmente a construção da Tabela 11, foram contabilizados os resultados individuais de cada MTEC utilizado, com a finalidade de comparar com o desempenho da utilização dos MTECs em conjunto. Estes resultados estão apresentados na Tabela 12.

Tabela 12 – Comparativo de desempenho de cada MTEC a cada 10 textos

Posição	Número de acertos do classificador	Número de acertos do KG	Número de acertos do WE
10	8	10	10
20	15	20	14
30	24	30	21
40	34	33	27
50	44	35	37
56	48	41	39
60	52	45	39
70	56	49	40
80	56	49	47
90	56	53	56
100	56	54	56
112	56	56	56

Fonte: autor

Considerando o resultado de todo o ranking (Apêndice B) é possível estabelecer a matriz confusão mostrada na Tabela 13 (a), (b) e (c) que estabelece os pontos de corte do ranking nas posições 25, 50 e 75.

Tabela 13 – Matriz confusão para o r*anking* final

(a) Ponto de corte em 25 posições

Matriz confusão: *Ranking* final		
Atual / Predito	Ideias	Projetos gerais
Ideias	25	0
Projetos gerais	31	56

(b) Ponto de corte em 50 posições

Matriz confusão: *Ranking* final		
Atual / Predito	Ideias	Projetos gerais
Ideias	47	3
Projetos gerais	8	53

(c) Ponto de corte em 75 posições

Matriz confusão: *Ranking* final		
Atual / Predito	Ideias	Projetos gerais
Ideias	54	21
Projetos gerais	2	35

Fonte: autor

A partir das matrizes confusão apresentadas, foram gerados os índices de precisão, acurácia, revocação, F1-score e *kappa* apresentados na Tabela 14

Tabela 14 - Índices do *ranking* final

	Indice	Valor
Ponto de corte em 25 posições	*Acurácia*	0,723
	Precision	0,446
	Recall	1,000
	F1-Score	0,617
	Kappa	0,446
Ponto de corte em 50 posições	*Acurácia*	0,900
	Precision	0,854
	Recall	0,940
	F1-Score	0,895
	Kappa	0,801
Ponto de corte em 75 posições	*Acurácia*	0,795
	Precision	0,964
	Recall	0,720
	F1-Score	0,824
	Kappa	0,589

Fonte: autor

A partir da Tabela 13 é possível observar que a melhor condição ocorre no ponto de corte igual a 50. Porém, isto não reflete uma condição de realidade, onde a precisão deveria

ser muito alta no ponto de corte de 25. Isto está ocorrendo pois a matriz confusão não é a melhor forma de analisar o *ranking*, visto que ela conta falsos positivos e verdadeiros negativos a partir de valores que não foram analisados, visto que um ponto de corte foi estabelecido. Neste sentido, a melhor forma de se analisar os resultados com base em um *ranking* é provido na tabela 11, a partir da identificação do total de acertos dentro de um grupo de ideias, ou seja, levando-se em conta um ponto de corte específico.

5.3 DISCUSSÃO DOS RESULTADOS

A partir dos resultados mostrados na Tabela 11 é possível verificar que o *ranking* fornece uma ordenação com alto índice de precisão. Em uma tarefa de seleção de ideias realizadas por um especialista, o *ranking* fornece informações importantes que podem ser utilizadas para a mineração de ideias e também para a seleção das mesmas por especialistas visando definir quais deveriam ser desenvolvidas.

A justificativa para a utilização da composição das 3 MTECs é apresentada pela Tabela 12 que mostra que os resultados individuais são inferiores aos resultados da composição dos MTECs no cálculo do *ranking* final.

Um ponto importante a ser observado é que foi realizado uma ordenação (*ranking*) a partir dos conjuntos de textos que representam ideias versus textos que representam projetos não executados. Isso foi estabelecido propositalmente para captar uma condição próxima a realidade. Pelo fato de serem conteúdos similares, mesmo os textos que representam projetos não executados podem também conter critérios de seleção de ideias, o que possibilita ainda mais corroborar o resultado final do modelo. Com este intuito, foram lidos os textos considerados falsos positivos (6) e constatou-se que todos continham a presença de critérios, conforme indicado pelo *ranking*.

Considerando os índices de avaliação obtidos, é importante realizar uma comparação com os principais trabalhos publicados na mesma área, considerando os últimos anos, os parágrafos a seguir descrevem os resultados obtidos.

O estudo de Ozcan *et al*. (2021) utilizou Mineração de Ideias (MI) sobre uma conjunto de dados de *tweets* para explorar tendências e recuperar ideias para vários fins, como desenvolvimento de produtos, tecnologia e sustentabilidade. As métricas obtidas usando 900 amostras de treinamento. Obtiveram a maior pontuação de *f1-score* de 0.564 utilizando SVM.

Os resultados mostram que a maior acurácia (78.4%) e precisão (0.526) foram obtidos através de SVM.

Já estudo de Alksher *et al.* (2018b) propõe uma nova abordagem léxico-sintática a ser utilizada em MI que enfatiza a caracterização da ideia existente. Esta abordagem modela as relações semânticas entre os atributos textuais que provavelmente compreendem ideias e excluem termos de relação que são considerados como ruídos para o modelo. O sistema foi construído em *Python*® utilizando os pacotes *nlkt*® e *Wordnet*®. A avaliação foi realizada com base em 3 cenários de estudo, onde os resultados da métrica precisão divulgados foram 0.758, 0.940 e 0.929.

Em Christensen *et al.* (2017b) foi realizado um estudo onde o principal objetivo consistia na investigação em como ideias são expressas e qual a natureza destas em comunidades *online*. Para chegar aos resultados, foram coletadas ideias de duas comunidades *online*. Uma comunidade relacionada ao interesse por cerveja e uma comunidade com interesses por brinquedos. Utilizou-se os mínimos quadrados parciais como método de classificação de texto sendo este comparado a um classificador tradicional, no caso SVM. Os mínimos quadrados parciais demonstraram 92% de *f1-score* no conjunto de dados sobre cerveja e 94% de *f1-score* no conjunto de dados sobre brinquedos. Já o SVM apresentou *f1-score* 93% no conjunto de dados sobre cerveja e 95% de *f1-score* no conjunto de dados sobre brinquedos. O autor afirma que seus resultados são interessantes pois confirmam que mínimos quadrados parciais deve ser considerado para tarefas futuras de classificação de texto, visto que estes permitem identificação de termos e variáveis importantes, conduzindo à classificação automática.

Comparando o modelo proposto nesta tese com os três trabalhos mais recentes em que se encontram resultados similares de acurácia, percebe-se que considerando as primeiras 40 posições do *ranking*, os resultados obtidos de 95% são próximos aos trabalhos publicados. Ao considerar o *ranking* com um corte na posição 56 (definido em função da quantidade de ideias do conjunto de teste que é de 112 ideias – definido na seção 3.4.4.2), o resultado obtido foi de 89%, o que também indica uma proximidade com os resultados atuais publicados.

Analisando os resultados dos graus de pertinência, as três últimas colunas da Tabela 10, é possível observar que existe alguma correspondência com os critérios obtidos na etapa de separação do conjunto de dados conforme critérios de especialistas, porém considerada baixa. De qualquer forma, foram contabilizados os acertos, considerando nesta contagem também os acertos parciais, visto que cada texto pode ser classificado por mais de um critério.

Exemplificando, caso uma ideias tenha sido inicialmente classificada por 2 critérios e a MTEC acertou 1 deles, considera-se então um acerto de 50%. O resultado obtido considerando o somatório de todos os acertos parciais e totais foi de 54%, o que pode ser melhorado. Porém, isto pode ser justificado pela falta de uma base de dados já classificada por critérios de especialistas.

A contribuição prática do trabalho como um todo consiste em auxiliar e simplificar a tarefa que um especialista possui na identificação e seleção de ideias a partir de um conjunto de dados de textos. Executando o modelo este seria capaz de receber um *ranking* e um grau de pertinência a qual(is) critério(s) cada ideia se enquadra. Neste sentido, o resultados obtidos demonstram que a apresentação final das ideias na forma de um *ranking* tende a facilitar e agilizar o trabalho de um especialista na identificação e seleção de ideias. A contribuição teórica está no fato de mostrar que os MTECs são capazes de modelar conhecimento humano e interconectar este conhecimento com resultados obtidos a partir do aprendizado de máquina na promoção de conteúdo que auxilia na tarefa de mineração de ideias. O fato dos MTECs terem sido combinados tornou possível uma maior explicitação em como os critérios e os diversos índices calculados impactam na definição da importância de cada ideia por meio de uma ordenação (*ranking*). Além disso, demonstrou que os resultados estão em linha e, em algumas situações, superiores aos os trabalhos mais atuais publicados. Todavia, percebe-se ainda que há espaço para melhorias, como a incorporação de novos MTECs e, também, a adoção de estratégias que possibilitem o treinamento contínuo após uma validação humana nos resultados do modelo.

5.4 CONSIDERAÇÕES FINAIS

O capítulo demonstrou inicialmente os resultados das avaliações das etapas do modelo no sentido de definir as melhores técnicas para a sua instanciação. Cada MTEC utilizado foi amplamente testado e, a partir disto, foram definidos os parâmetros de funcionamento e a melhor configuração de utilização.

Na sequência o modelo foi efetivamente avaliado e seus resultados apresentados através de *ranking* que promove uma ordenação do conjunto de dados de forma que as ideias, considerando um conjunto de dados com os tipos de textos, fiquem evidenciadas nas primeiras posições. Por fim, os resultados foram analisados frente aos trabalhos na área de MI estando em linha com resultados destes trabalhos recentes quanto a acurácia na identificação

de ideias. Todavia, cabe mencionar que nenhum destes trabalhos promove uma visão de *ranking* com o intuito de auxiliar especialistas na identificação e seleção de potencias ideias para implementação. Por fim, foram apresentadas as principais contribuições práticas e teóricas do modelo.

6 CONCLUSÕES

As ideias impulsionam o desenvolvimento de novos produtos e são informações estratégicas nas organizações. Tal afirmação reside no fato de que as organizações procuram divulgar apenas internamente, não disponibilizando mais suas bases de ideias na *web*. Por outro lado, as organizações utilizam-se de Mineração de Ideias (MI) para automatizar seus processos e torná-las mais competitivas. Porém, mesmo com o avanço da MI ainda existem alguns desafios, como as pesquisas se concentrarem em explorar algoritmos para extrair padrões a partir de textos, mas ignorarem as estruturas de conhecimento existentes nos textos. Outro desafio reside no fato de que o conhecimento do especialista é fundamental na etapa de seleção de ideias, existindo uma necessidade de traduzir esse conhecimento de domínio dos especialistas humanos em uma estrutura de aprendizagem coesa e expressiva.

Diante disto, esta tese foi formulada para suprir as lacunas de conhecimento acima mencionadas e contribuir em como um modelo de MI pode agregar tais características. Para tal, foram traçados os objetivos geral e específicos, onde nos parágrafos a seguir se discute se os mesmos foram ou não atingidos.

Considerando o desenvolvimento deste trabalho, que possui como objetivo geral a proposição de um modelo voltado à identificação de ideias a partir de fontes de informação não estruturadas levando-se em conta critérios de escolha de ideias utilizados por especialistas, pode-se afirmar a partir dos resultados o mesmo foi cumprido. Todas os MTECs utilizados neste modelo foram treinados para identificação dos critérios dos especialistas e apresentaram um aprimoramento em relação ao seu uso tradicional na identificação de deias. A composição dos resultados individuais, onde é levado em consideração a quantidade de critérios identificados, faz com que os resultados finais sejam mais precisos e similares ao processo que ocorre na tarefa de identificação e seleção de ideias realizada por um especialista humano.

Quanto aos objetivos específicos também pode-se dizer que foram todos alcançados. Para elencar os principais critérios utilizados por especialistas na identificação de ideias foi realizada uma revisão da literatura. Par atingir o objetivo específico de selecionar os principais métodos e técnicas de mineração de ideias que pudessem promover suporte ao desenvolvimento do modelo proposto considerando os critérios utilizados por especialistas, foram utilizados classificadores de texto e estruturas de conhecimento, mais especificamente os *Knowledge Graphs* e os *Word Embeddings*. Para estes métodos e técnicas foram

realizados diversos testes visando obter as melhores configurações que que pudessem atingir resultados adequados durante a avaliação do modelo proposto.

Quanto ao objetivo específico de apresentar as ideias identificadas em formato de *ranking* explicitando a aderência das mesmas aos critérios utilizados pelos especialistas, foram demonstrados os cálculos utilizados para a composição da ordenação. O *ranking* final (apresentado no Apêndice A) apresenta todas as variáveis e os resultados dos cálculos, onde a aderência a cada critério é apresentada em formato de porcentagem.

E, para finalizar os objetivos específicos, visando evidenciar a viabilidade do modelo proposto através do desenvolvimento de um protótipo e aplicação deste em cenários de estudo, foram elaborados dois cenários de estudo. O primeiro é composto por um conjunto de dados dividido em textos que representam ideias e textos comuns, enquanto o segundo cenário se utiliza de um conjunto de dados dividido em textos que representam ideias e textos que representam projetos não executados. Esta divisão se fez necessária pois o treinamento do modelo exige uma separação mais precisa dos dados, enquanto a aplicação final exige uma condição mais próxima a realidade. O primeiro cenário é destinado a avaliação e o segundo aos testes efetivos do modelo proposto.

Os resultados do primeiro cenário de estudo mostraram que o classificador indicado para este modelo é o *Naive Bayes.* Já os *Knowledge Graphs* forneceram os conceitos e as relações de cada critério, enquanto os *Word Embeddings* forneceram um centroide geral de cada critério, que quando comparado com o centroide das 5 principais palavras de cada texto, permitem captar a aderência a este critério através de uma distância entre estes pontos. Todos estes resultados (classificadores, KGs e WEs) foram aplicados ao segundo cenário de estudos que, por sua vez, permitiu a construção de uma relação de ideias de maneira ordenada por meio de um índice que representa o quanto determinado texto se caracteriza como uma ideia.

Para evidenciar a viabilidade e analisar a efetividade do modelo foram combinados *softwares* de terceiros com o desenvolvimento de um protótipo permitindo uma instanciação do modelo. Desta forma, como saída do modelo é apresentada uma relação que determina a importância de cada texto, seja uma ideia ou texto comum, em formato de *ranking*. Este *ranking* objetiva ao fim, auxiliar especialistas durante a avaliação e identificação de ideias que possuam potencial de implementação. Os resultados são promissores, constituindo um modelo que pode ser aplicado na prática. Estes resultados foram discutidos comparando-se com trabalhos de MI atuais constatando-se similaridades e, por vezes, superioridade. Nestes trabalhos não foi identificada a proposição de um *ranking* que permitisse ao especialista

identificar claramente a importância de cada ideia. É relevante registrar que o modelo, apesar de ter demonstrado potencial para aplicação prática, possui espaço para melhorias, onde novos MTECs podem ser adicionados na composição dos resultados.

Contudo, algumas limitações podem ser elencadas no modelo proposto. Entre elas citam-se a necessidade de realização de testes com bases de dados maiores, até mesmo promovendo uma validação estatística. Apesar disso, cita-se que por razões organizacionais estratégicas, atualmente as bases de ideias têm se tornando restritas, dificultando o avanço em testes mais aprimorados. Outra limitação está no fato do modelo trabalhar somente no idioma português, sendo interessante abrir possibilidade para novos idiomas.

Diante de todos os detalhamentos dos resultados obtidos neste trabalho, pode-se concluir que a pergunta de pesquisa formulada *"Como identificar ideias a partir de fontes de dados não estruturados, considerando critérios de escolha utilizados por especialistas durante o processo de seleção de ideias?"* foi respondida. Os resultados demonstram que é possível identificar ideias a partir de fontes de dados não estruturadas considerando critérios de escolha utilizados por especialistas durante o processo de seleção de ideias. Ademais, pode-se afirmar que a promoção do resultado na forma de uma lista ordenada (*ranking*) constitui-se em uma característica importante e diferenciada do modelo, que pode facilitar a tarefa de identificação e seleção de ideias com potencial de implementação por especialistas.

6.1 TRABALHOS FUTUROS

Durante o desenvolvimento desta tese outras possibilidades de implementação do modelo foram consideradas.

Primeiramente, pode ser citada a possibilidade de inserção de novos critérios de especialistas. A ampliação da capacidade de identificação de outros critérios de maneira mais automatizada promoveria a utilização em outros domínios, com critérios diferentes dos inicialmente propostos.

A inserção de novos MTECs como redes neurais, ontologias, entre outros, também poderiam ser considerados na evolução do modelo. A título de observação, durante a fase final do desenvolvimento da tese, mais especificamente, após a etapa de qualificação, chegou-se a testar redes neurais convolucionais (do inglês Convolutional Neural Network – CNN). Os resultados iniciais não se apresentaram satisfatórios na classificação de textos como ideias e

não ideias. Todavia, novos testes e a correta identificação dos hiperparâmetros de uma CNN com este intuito pode contribuir para a evolução do modelo aqui proposto.

Outro ponto a ser destacado reside na incorporação de uma estratégia de treinamento contínuo, de modo que ideias selecionadas por especialistas na etapa final de avaliação pudessem ser incorporadas ao modelo de treinamento. Desta forma, objetiva-se um aprendizado incremental e recorrente, possibilitando a evolução e aprimoramento dos resultados.

Por fim, mas sem exaurir as possibilidades, cita-se a criação de uma interface amigável integrando todos os MTECs utilizados na confecção do modelo. A integração de todos os aplicativos utilizados através de uma interface poderia viabilizar um sistema com potencial acadêmico e comercial.

REFERÊNCIAS

ABABNEH, J. Application of Naïve Bayes, Decision Tree, and K-Nearest Neighbors for Automated Text Classification. **Modern Applied Science**, v. 13, n. 11, p. 31, 2019.

AGGARWAL, C. C.; ZHAI, C. X. A survey of text classification algorithms. **Mining Text Data**, v. 9781461432, p. 163–222, 2012.

AGUIAR, C. Informação e atividades de desenvolvimento científico, tecnológico e industrial: tipologia proposta com base em análise funcional. **Ciência da Informação**, v. 47, n. 1, p. 7–15, 1991.

ALIYEVA, D. et al. Combining Dual Word Embeddings with Open Directory Project Based Text Classification. **Proceedings of 2018 IEEE 17th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2018**, p. 179–186, 2018.

ALKSHER, M. et al. A review of methods for mining idea from text. **2016 3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016 - Conference Proceedings**, p. 88–93, 2016.

ALKSHER, M. et al. **A Framework for Idea Mining Evaluation**. 16th International Conference on New Trends in Intelligent Software Methodology Tools, and Techniques, SoMeT 2017. **Anais**...2017

ALKSHER, M. et al. Feasibility of Using the Position as Feature for Idea Identification from Text. **2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)**, p. 1–6, 2018a.

ALKSHER, M. et al. Effective Idea Mining Technique Based on Modeling Lexical Semantic. v. 96, n. 16, p. 5350–5362, 2018b.

ALLAHYARI, M. et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. **KDD Bigdas**, 2017.

ALMEIDA, E. C. E. DE. **O portal de periódicos da Capes: estudo sobre a sua evolução e utilização**. [s.l.] Universidade de Brasília, 2006.

ALTINEL, B.; GANIZ, M. C. Semantic text classification: A survey of past and recent advances. **Information Processing and Management**, v. 54, n. 6, p. 1129–1153, 2018.

AMABILE, T. M. **Creativity in context: The social psychology of creativityBoulder, CO: Westview**, 1996.

AUBAID, A. M.; MISHRA, A. Text classification using word embedding in Rule-based

methodologies: A systematic mapping. **TEM Journal**, v. 7, n. 4, p. 902–914, 2018.

AYELE, W. Y. Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset. **International Journal of Advanced Computer Science and Applications**, v. 11, n. 6, p. 20–32, 2020.

AYELE, W. Y.; JUELL-SKIELSE, G. A Process Model for Generating and Evaluating Ideas: The Use of Machine Learning and Visual Analytics to Support Idea Mining. **International Conference on Electronic Government and the Information SystemsPerspective**, v. 12394 LNCS, p. 189–203, 2020.

AYELE, W. Y.; JUELL-SKIELSE, G. A Systematic Literature Review about Idea Mining : The Use of Machine-Driven Analytics to Generate Ideas. **Advances in Intelligent Systems and Computing**, v. 1364, p. 744–762, 2021.

AZMAN, A. et al. Optimization of idea mining model based on text position weight. **International Journal of Advanced Trends in Computer Science and Engineering**, v. 8, n. 1.4 S1, p. 120–125, 2019.

AZMAN, A. et al. A Framework for Automatic Analysis of Essays Based on Idea Mining. **Lecture Notes in Electrical Engineering**, v. 603, p. 639–648, 2020.

BADAWI, D.; ALTINÇAY, H. A novel framework for termset selection and weighting in binary text classification. **Engineering Applications of Artificial Intelligence**, v. 35, p. 38–53, 2014.

BAYOUDE, K. et al. How machine learning potentials are transforming the practice of digital marketing: State of the art. **Periodicals of Engineering and Natural Sciences**, v. 6, n. 2, p. 373–379, 2018.

BOLLACKER, K. et al. Freebase: A collaboratively created graph database for structuring human knowledge. **Proceedings of the ACM SIGMOD International Conference on Management of Data**, p. 1247–1249, 2008.

BRASIL, G. F. DO. **Portal Brasileiro de Dados Abertos**. Disponível em: <https://dados.gov.br/>. Acesso em: 29 set. 2021.

BREIMAN, L. Randon Forests. **Machinelearning202.Pbworks.Com**, p. 1–35, 1999.

BRINDHA, S.; PRABHA, K.; SUKUMARAN, S. A survey on classification techniques for text mining. **ICACCS 2016 - 3rd International Conference on Advanced Computing and Communication Systems: Bringing to the Table, Futuristic Technologies from Arround the Globe**, v. 01, n. i, p. 1–5, 2016.

BURRELL, G.; MORGAN, G. **Social paradigms and organisational analysis: elements of**

the sociology of corporate life. [s.l.] Books, Great Britain: Heinemann Education, 1979.

BUTNARU, A.; IONESCU, R. T. UnibucKernel: A kernel-based learning method for complex word identification. n. April, p. 175–183, 2018.

BUTNARU, A. M.; IONESCU, R. T. From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings. **Procedia Computer Science**, v. 112, p. 1783–1792, 2017.

CAMPBELL, J. B.; WYNNE, R. H. **Introduction to remote sensing**Guilford Press, , 2011.

CHEN, J. et al. Naive bayes with correlation factor for text classification problem. **Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019**, p. 1051–1056, 2019.

CHEN, X.; CARDIE, C. Multinomial adversarial networks for multi-domain text classification. **NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference**, v. 1, p. 1226–1240, 2018.

CHEN, X.; JIA, S.; XIANG, Y. A review: Knowledge reasoning over knowledge graph. **Expert Systems with Applications**, v. 141, 2020.

CHRISTENSEN, K. et al. In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining. **Creativity and Innovation Management**, v. 26, n. 1, p. 17–30, 2017a.

CHRISTENSEN, K. et al. Mining online community data: The nature of ideas in online communities. **Food Quality and Preference**, v. 62, n. December 2016, p. 246–256, 2017b.

CLARKE, M.; HORTON, R. Bringing it all together: Lancet-Cochrane collaborate on systematic reviews. **The Lancet**, v. 357, n. 9270, p. 1728, 2001.

COOK, D. J.; MULROW, C. D.; HAYNES, R. B. Systematic reviews: synthesis of best evidence for clinical decisions. **Annals of internal medicine**, v. 126, n. 5, p. 376–80, 1997.

COOPER, B. R.; EDGETT, S. Ideation for Product Innovation : What are. **Development**, n. March 2008, p. 9, 2008.

COSTA, A. J. S. DA et al. Desenvolvimento De Ferramenta Para Gestao De Ideias No Modelo De Inovacao Mgpdi. **Proceedings of the 17th CONTECSI International Conference on Information Systems and Technology Management**, v. 1, p. 34223–34245, 2021.

COUSSEMENT, K.; VAN DEN POEL, D. Churn prediction in subscription services: An

application of support vector machines while comparing two parameter-selection techniques. **Expert Systems with Applications**, v. 34, n. 1, p. 313–327, 2008.

CUNNINGHAM, E. et al. Using text classification methods to detect malware. **CEUR Workshop Proceedings**, v. 2563, p. 95–103, 2019.

CUPANI, A. **Filosofia da Tecnologia: um convite**. Florianópolis: Editora da UFSC, 2011.

DE RAADT, A. et al. Kappa Coefficients for Missing Data. **Educational and Psychological Measurement**, v. 79, n. 3, p. 558–576, 2019.

DEMSAR, J. et al. Orange: Data Mining Toolbox in Python Janez. **International Journal of Conservation Science**, v. 7, n. SpecialIssue1, p. 295–300, 2016.

DENG, X. et al. Feature Selection for Text Classification. p. 257–276, 2019.

DHAVALIKAR, A. S.; CHOUDHARI, P. C. Classification of Oil Spills and Look-Alikes from Sar Images Using Bag of Visual Words Method of Feature Extraction. **2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS**, p. 3428–3431, 2021.

DÍAZ-GARCÍA, C.; GONZÁLEZ-MORENO, Á.; SÁEZ-MARTÍNEZ, F. J. Eco-innovation: insights from a literature review.: EBSCOhost. **Organization & Management Volume**, v. 17, n. 1, p. 1–5, 2015.

DISSELKAMP, M. **Innovationsmanagement: Instrumente und Methoden zur Umsetzung im Unternehmen**. [s.l: s.n.].

DWIVEDI, S. K.; ARYA, C. Automatic Text Classification in Information retrieval. p. 1–6, 2016.

EGC-UFSC. **EGC - Áreas de Concentração**. Disponível em: <http://www.egc.ufsc.br/pos-graduacao/programa/areas-de-concentracao/>. Acesso em: 20 mar. 2019.

ELEKES, A. et al. Learning from Few Samples: Lexical Substitution with Word Embeddings for Short Text Classification. p. 111–119, 2019.

ELERUD-TRYDE, A.; HOOGE, S. Beyond the generation of ideas: Virtual idea campaigns to spur creativity and innovation. **Creativity and Innovation Management**, v. 23, n. 3, p. 290–302, 2014.

ELHASSAN, R.; ALI, M. The Impact of Feature Selection Methods for Classifying Arabic Texts. **2nd International Conference on Computer Applications and Information Security, ICCAIS 2019**, 2019.

FELDMAN, R.; DAGAN, I. Knowledge Discovery in Textual Databases (KDT). **International Conference on Knowledge Discovery and Data Mining (KDD)**, p. 112–117, 1995.

FERIOLI, M. et al. Evaluation of the Potential Performance of Innovative Concepts in the Early Stages of the New-Product Development Process (Npdp). **Design**, p. 1139–1148, 2008.

FLYNN, M. et al. Idea Management for Organisational Innovation. **International Journal of Innovation Management**, v. 07, n. 04, p. 417–442, 2003.

FRIZZARINI, C.; LAURETTO, M. S. Proposta de um Algoritmo para Indução de Árvores de Classificação para Dados Desbalanceados. **IX SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO**, v. IX, p. 722–733, 2013.

GHIMIRE, B. et al. An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. **GIScience and Remote Sensing**, v. 49, n. 5, p. 623–643, 2012.

GOLDBERG, Y. Neural Network Methods for Natural Language Processing. **Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International**, v. 37, p. 0–2, 2013.

GONÇALVES, A. L. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à Engenharia e Gestão do Conhecimento**. [s.l: s.n.].

GRÜNLING, N. Fuzzy front-end of Entrepreneurship Developing a Business Idea Selection Framework. **Nova School of Business and Economics**, 2017.

GUO, B. et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. **Neurocomputing**, v. 363, p. 366–374, 2019.

GUO, Y. et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes. **Proceedings of the 2010 Workshop on Biomedical Natural Language Processing**, p. 99–107, 2010.

HA, S.; GEUM, Y. Identifying new innovative services using M&A data: An integrated approach of data-driven morphological analysis. **Technological Forecasting and Social Change**, v. 174, n. February 2021, p. 121197, 2022.

HARRIS, Z. S. Distributional Structure. **Word**, v. 10, n. 2–3, p. 146–162, 1954.

HEIMERL, F.; GLEICHER, M. Interactive Analysis of Word Vector Embeddings. **Computer Graphics Forum**, v. 37, n. 3, p. 253–265, 2018.

HO, T. K. Random Decision Forests Tin Kam Ho Perceptron training. **Proceedings of 3rd International Conference on Document Analysis and Recognition**, p. 278–282, 1995.

HOAI NAM, L. N.; QUOC, H. B. Integrating Low-rank Approximation and Word Embedding for Feature Transformation in the High-dimensional Text Classification. **Procedia Computer Science**, v. 112, p. 437–446, 2017.

HOTHO, A. et al. A Brief Survey of Text Mining. p. 1–37, 2005.

HOU, R. et al. Unstructured big data analysis algorithm and simulation of Internet of Things based on machine learning. **Neural Computing and Applications**, v. 32, n. 10, p. 5399–5407, 2020.

IBRAHIM, K. R.; GILMOUR, R. F. Method to Identify Quality Ideas for New Product Development : RUOG & ODVV 3OD \ HU. **PICMET '16**, p. 2515–2531, 2016.

ISLAM, M. Z. et al. A semantics aware random forest for text classification. **International Conference on Information and Knowledge Management, Proceedings**, p. 1061–1070, 2019.

JIANG, S.; ZHAI, C.; MEI, Q. Exploiting Knowledge Graph to Improve Text-based Prediction. **Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018**, p. 1407–1416, 2018.

JIANG, X. et al. BaKGraSTeC: A background knowledge graph based method for short text classification. **Proceedings - 11th IEEE International Conference on Knowledge Graph, ICKG 2020**, p. 360–366, 2020.

JOACHIMS, T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. **Fourteenth International Conference on Machine Learning**, p. 143–151, 1997.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 1398, p. 137–142, 1998.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **SCIENCE sciencemag**, v. 349, n. 6245, 2015.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An introduction to natural language processing**. [s.l: s.n.].

KADHIM, A. I. Survey on supervised machine learning techniques for automatic text classification. **Artificial Intelligence Review**, v. 52, n. 1, p. 273–292, 2019.

KARAMI, A. et al. Twitter and Research: A Systematic Literature Review through Text Mining. **IEEE Access**, v. 8, p. 67698–67717, 2020.

KARIMI-MAJD, A. M.; MAHOOTCHI, M. A new data mining methodology for generating new service ideas. **Information Systems and e-Business Management**, v. 13, n. 3, p. 421–443, 2015.

KHAN, K. et al. Mining opinion components from unstructured reviews: A review. **Journal of King Saud University - Computer and Information Sciences**, v. 26, n. 3, p. 258–275, 2014.

KILIMCI, Z. H.; AKYOKUS, S. Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification. **Complexity**, v. 2018, 2018.

KIM, J.; MACDUFFIE, J. P.; PIL, F. K. Employee voice and organizational performance: Team versus representative influence. **Human Relations**, v. 63, n. 3, p. 371–394, 2010.

KIM, J.; PARK, Y. Leveraging ideas from user innovation communities: using text-mining and case-based reasoning. **R and D Management**, v. 49, n. 2, p. 155–167, 2019.

KISHIDA, K. Technical issues of cross-language information retrieval: A review. **Information Processing and Management**, v. 41, n. 3, p. 433–455, 2005.

KLEIN, M.; GARCIA, A. C. B. High-speed idea filtering with the bag of lemons. **Decision Support Systems**, v. 78, p. 39–50, 2015.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. **International Joint Conference of Artificial Intelligence**, n. March 2001, 1995.

KOWSARI, K. et al. Text classification algorithms: A survey. **Information (Switzerland)**, v. 10, n. 4, p. 1–68, 2019.

KUMAR, R.; KAUR, J. Random forest-based sarcastic tweet classification using multiple feature collection. **Intelligent Systems Reference Library**, v. 163, p. 131–160, 2020.

LANDIS, J. R.; KOCH, G. G. The Measurement of Observer Agreement for Categorical Data. **Biometrics**, v. 33, n. 1, p. 159, 1977.

LEE, A. V. Y.; TAN, S. C. Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. v. 4, n. 2107, p. 76–101, 2017a.

LEE, A. V. Y.; TAN, S. C. Discovering Dynamics of an Idea Pipeline : Understanding Idea Development within a Knowledge Building Discourse. p. 119–128, 2017b.

LI, S. M. et al. New product idea selection in the fuzzy front end of innovation: A fuzzy best-worst method and group decision-making process. **Mathematics**, v. 9, n. 4, p. 1–18, 2021.

LI, X. L.; LIU, B.; NG, S. K. Negative training data can be harmful to text classification. **EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference**, n. October, p. 218–228, 2010.

LI, X.; LI, L.; CHEN, Z. Toward Extenics-Based Innovation Model on Intelligent Knowledge Management. **Annals of Data Science**, v. 1, n. 1, p. 127–148, 2014.

LIN, Y. et al. Learning entity and relation embeddings for knowledge graph completion. **Proceedings of the National Conference on Artificial Intelligence**, v. 3, p. 2181–2187, 2015.

LIU, H.; GOULDING, J.; BRAILSFORD, T. Towards computation of novel ideas from corpora of scientific text. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 9285, p. 541–556, 2015.

LOYOLA-GONZÁLEZ, O.; MEDINA-PÉREZ, M. A.; CHOO, K. K. R. A Review of Supervised Classification based on Contrast Patterns: Applications, Trends, and Challenges. **Journal of Grid Computing**, v. 18, n. 4, p. 797–845, 2020.

MAGNUSSON, P. R.; WÄSTLUND, E.; NETZ, J. Exploring Users' Appropriateness as a Proxy for Experts When Screening New Product/Service Ideas. **Journal of Product Innovation Management**, v. 33, n. 1, p. 4–18, 2014.

MARIN, A. et al. Learning phrase patterns for text classification using a knowledge graph and unlabeled data. **Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH**, n. September, p. 253–257, 2014.

MAXWELL, A. E.; WARNER, T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: An applied review. **International Journal of Remote Sensing**, v. 39, n. 9, p. 2784–2817, 2018.

MAYFIELD, E.; ROSÉ, C. P. **LightSIDE: Text Mining and Machine Learning User ' s Manual**. [s.l: s.n.].

MCCALLUM, A.; NIGAM, K. Employing EM and Pool-Based Active Learning for Text Classi cation 2 Probabilistic Framework for Text Classi cation. 1998.

MENG, Y.; WANG, G.; LIU, Q. Multi-layer convolutional neural network model based on prior knowledge of knowledge graph for text classification. **2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019**, p. 618–624, 2019.

MIKOLOV, T. Learning Representations of Text using Neural Networks. **NIPS Deep Learning Workshop**, p. 1–31, 2013.

MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. p. 1–12, 2013.

MORAIS, M. DE O.; MARIA, D. F.; OLIVEIRA, L. M. DE. A Inovação e a Indústria 4.0: Proposta para utilização de elementos para uma organização competitiva. **Research, Society and Development**, v. 10, n. 8, p. e51210817685, 2021.

MOREO, A.; ESULI, A.; SEBASTIANI, F. **Word-class embeddings for multiclass text classification**. [s.l.] Springer US, 2021. v. 35

MORGAN, G. **Paradigms, Metaphors, and Puzzle Solving in Organization Theory**. [s.l.] Cornell University, 1980.

MOUSAVI, S. M.; TORABI, S. A.; TAVAKKOLI-MOGHADDAM, R. A Hierarchical Group Decision-Making Approach for New Product Selection in a Fuzzy Environment. **Arabian Journal for Science and Engineering**, v. 38, n. 11, p. 3233–3248, 2013.

NAILI, M.; CHAIBI, A. H.; BEN GHEZALA, H. H. Comparative study of word embedding methods in topic segmentation. **Procedia Computer Science**, v. 112, p. 340–349, 2017.

NILC. **Repositório de Word Embeddings do NILC**. Disponível em: <http://www.nilc.icmc.usp.br/embeddings>.

NISULA, A. M.; KIANTO, A. Evaluating and developing innovation capabilities with a structured method. **Interdisciplinary Journal of Information, Knowledge, and Management**, v. 8, p. 59–82, 2013.

NONAKA, IKUJIRO; TAKEUCHI, H. The best Japanese companies offer aguide to the organizational roles, structures, and practices that produce consinuous innovarion. In: **The Knowlege-Creating Company**. [s.l: s.n.]. v. 103p. 411.

OCDE. Manual de Frascati 2002: metodologia proposta para definição da pesquisa e desenvolvimento experimental. p. 324, 2013.

OLSSON, E. H.; LANDSTRÖM, K. **Enhancing the Innovation Performance by Employing Criteria**. [s.l.] University of Gothenburg, 2020.

OZCAN, S. et al. Social media mining for ideation: Identification of sustainable solutions and opinions. **Technovation**, v. 107, n. April 2020, p. 102322, 2021.

OZER, M. Managing the selection process for new product ideas. **Research Technology Management**, v. 47, n. 4, p. 11, 2004.

PACHECO, R. C. DOS S.; TOSTA, K.; FREIRE, P. DE S. Interdisciplinaridade vista como um processo complexo de construção do conhecimento : uma análise do Programa de Pós-Graduação EGC / UFSC. **Knowledge Management**, v. 7, n. 12, p. 136–159, 2010.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Machine learning and soil sciences: A review aided by machine learning tools. **Soil**, v. 6, n. 1, p. 35–52, 2020.

PAN, C. et al. Few-Shot Transfer Learning for Text Classification with Lightweight Word Embedding Based Models. **IEEE Access**, v. 7, p. 53296–53304, 2019.

PARCHETA, Z. et al. Combining Embeddings of Input Data for Text Classification. **Neural Processing Letters**, 2020.

PAUKKERI, M. Framework for Analyzing and Clustering Short Message. n. September, p. 239–247, 2009.

PAULHEIM, H. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. **Semantic Web**, v. 8, n. 3, p. 489–508, 2016.

PEFFERS, K. et al. A Design Science Research Methodology for Information Systems Research. **Journal of Management Information Systems**, v. 24, p. 45–78, 2007.

PEJIC-BACH, M. et al. Text mining of industry 4.0 job advertisements. **International Journal of Information Management**, v. 50, n. August 2019, p. 416–431, 2020.

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

PHU, V. N. et al. A decision tree using ID3 algorithm for English semantic analysis. **International Journal of Speech Technology**, v. 20, n. 3, p. 593–613, 2017.

POZZO, R. **Naturauffassungen in Philosophie, Wissenschaft**. vol4. ed. [s.l: s.n.].

QUINLAN, J. R. Induction of Decision Trees. **Research and Development in Expert Systems XV**, v. 1, n. Chapter 2, p. 15–26, 1986.

RANGANATHAN, J.; IRUDAYARAJ, A. S.; TZACHEVA, A. A. Automatic Detection of Emotions in Twitter Data - A Scalable Decision Tree Classification Method. n. July, 2018.

REATEGUI, E. et al. Sobek: a Text Mining Tool for Educational Applications. **Proceedings International Conference on Data Mining (DMIN)**, p. 59–64, 2011.

REZAEI, J. Best-worst multi-criteria decision-making method. **Omega (United Kingdom)**,

v. 53, p. 49–57, 2015.

RISH, I. An empirical study of the naive Bayes classifier. **IJCAI 2001 workshop on empirical methods in artificial intelligence**, v. 3, p. 41–46, 2001.

RÖLTGEN, A. T. et al. Development, implementation and evaluation of a digital idea management system. A case analysis. **Gruppe. Interaktion. Organisation. Zeitschrift fur Angewandte Organisationspsychologie**, v. 51, n. 1, p. 49–58, 2020.

ROTMENSCH, M. et al. Learning a Health Knowledge Graph from Electronic Medical Records. **Scientific Reports**, v. 7, n. 1, p. 1–11, 2017.

SADRIEV, A. R.; PRATCHENKO, O. V. Idea Management in the System of Innovative Management. **Mediterranean Journal of Social Sciences**, v. 5, n. 12, p. 155–158, 2014.

SAFAVIAN, R.; LANDGREBE, D. A Survey of Decision Tree Classifier. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 21, n. 3, p. 660–674, 1991.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 44, n. 1.2, p. 206–226, 2000.

SARIGIANNI, C. et al. **Innovation contests: How to design for successful idea selectionProceedings of the Annual Hawaii International Conference on System Sciences**, 2020.

SCHAFFER, C. Technical Note: Selecting a Classification Method by Cross-Validation. **Machine Learning**, v. 13, n. 1, p. 135–143, 1993.

SEBASTIANI, F. Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, v. 34, 2002.

SÉRGIO, M. C.; GONÇALVES, A. L. Analysis and interpretation of ideas: Proposal of a model. **Perspectivas em Ciencia da Informacao**, v. 24, n. 2, p. 54–71, 2019.

SÉRGIO, M. C.; SOUZA, J. A.; GONÇALVES, A. L. Idea Identification Model to Support Decision Making. **IEEE Latin America Transactions**, v. 15, p. 968–973, 2017.

SHUANG, K. et al. Convolution–deconvolution word embedding: An end-to-end multi-prototype fusion embedding method for natural language processing. **Information Fusion**, v. 53, n. May 2019, p. 112–122, 2020.

SILVA, A.; GOMBOLAY, M. Neural-encoding Human Experts' Domain Knowledge to Warm Start Reinforcement Learning. 2019.

SILVA, E. L. DA; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: UFSC, 2005.

SILVA, E. C.; PEDRON, C. D. Elementos Determinantes Para a Capacidade De Inovação Das Empresas: Uma Revisão Sistemática Da Literatura. **Revista Brasileira de Gestão e Inovação**, v. 7, n. 1, p. 45–63, 2019.

SINAPSE DA INOVAÇÃO. **Portal Sinapse da Inovação**. Disponível em: <http://sc.sinapsedainovacao.com.br/>. Acesso em: 5 jun. 2019.

SINGHAL, A. Modern Information Retrieval: A Brief Overview. **Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**, v. 24, n. 4, p. 35–43, 2001.

SINOARA, R. A. et al. Knowledge-enhanced document embeddings for text classification. **Knowledge-Based Systems**, v. 163, p. 955–971, 2019.

SONG, X.; SRIMANI, P. K.; WANG, J. Z. Hwe: Hybrid word embeddings for text classification. **ACM International Conference Proceeding Series**, p. 25–29, 2019.

STEIN, R. A.; JAQUES, P. A.; VALIATI, J. F. An analysis of hierarchical text classification using word embeddings. **Information Sciences**, v. 471, p. 216–232, 2019.

TAVARES, L. G.; LOPES, H. S.; LIMA, C. R. E. Estudo comparativo de métodos de aprendizado de máquina na detecçaoo de regiões promotoras de genes de escherichia coli. **Anais do I Simpósio Brasileiro de Inteligência Computacional.**, p. 8–11, 2007.

THORLEUCHTER, D.; HERBERZ, S.; POEL, D. VAN DEN. Mining social behavior ideas of Przewalski horses. **Lecture Notes in Electrical Engineering**, v. 121 LNEE, p. 649–656, 2011.

THORLEUCHTER, D.; VAN DEN POEL, D. Extraction of ideas from microsystems technology. **Advances in Intelligent and Soft Computing**, v. 168 AISC, n. VOL. 1, p. 563–568, 2012.

THORLEUCHTER, D.; VAN DEN POEL, D. Web mining based extraction of problem solution ideas. **Expert Systems with Applications**, v. 40, n. 10, p. 3961–3969, 2013.

THORLEUCHTER, D.; VAN DEN POEL, D. Idea mining for web-based weak signal detection. **Futures**, v. 66, p. 25–34, 2015.

THORLEUCHTER, D.; VAN DEN POEL, D. Identification of interdisciplinary ideas. **Information Processing and Management**, v. 52, n. 6, p. 1074–1085, 2016.

THORLEUCHTER, D.; VAN DEN POEL, D.; PRINZIE, A. Mining ideas from textual information. **Expert Systems with Applications**, v. 37, n. 10, p. 7182–7188, 2010.

TOPAZ, M. et al. Identifying Diabetes in Clinical Notes in Hebrew: A Novel Text Classification Approach Based on Word Embedding. **Studies in health technology and informatics**, v. 264, p. 393–397, 2019.

TOSI, M. D. L.; DOS REIS, J. C. SciKGraph: A knowledge graph approach to structure a scientific field. **Journal of Informetrics**, v. 15, n. 1, p. 101109, 2021.

TREVISAN, L. C.; PELOGIA, I.; DAMIAN, M. Gestão do conhecimento: diretrizes e práticas recomendadas às organizações. **Ciência da Informação**, v. 47, n. 2, p. 21–34, 2018.

TRIPATHY, A. et al. Extracting new product ideas from consumer blogs. **Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012**, n. 2010, p. 1–6, 2012.

VALDATI, A. D. B.; DANDOLINI, G. A. Critérios para seleção de ideias no front end da inovação ideas selection criteria in the innovation front end criterios de selección de ideas en el front end de la innovación. **Revista Eletrônica de Estratégia & Negócios**, 2019.

VALDATI, A. DE B. **Processo de Seleção de Ideias em Empresas Inovadoras**. [s.l.] Universidade Federal de Santa Catarina, 2017.

VAN DEN ENDE, J.; FREDERIKSEN, L.; PRENCIPE, A. The front end of innovation: Organizing search for ideas. **Journal of Product Innovation Management**, v. 32, n. 4, p. 482–487, 2015.

VAN ROSSUM, G. Python tutorial, Technical Report CS-R9526. **Centrum voor Wiskunde en Informatica (CWI)**, 1995.

VIEIRA, T. A TRANSFORMAÇÃO DIGITAL SOB A ÓTICA DA ENGENHARIA DO CONHECIMENTO: UMA REVISÃO SOBRE O USO DE ONTOLOGIAS COMO MODELO Thaianne Vieira 1. **X Congresso Internacional de Conocimiento e Innovacion**, 2020.

VIJAYAN, V. K.; BINDU, K. R.; PARAMESWARAN, L. A Comprehensive Study of Text Classification Algorithms. p. 1109–1113, 2017.

WANG, H.; GUO, K.; LIU, Z. Mixed word embedding method based on knowledge graph augment for text classification. **Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019**, p. 1618–1623, 2019.

WANG, Q. et al. Knowledge graph embedding: A survey of approaches and applications. **IEEE Transactions on Knowledge and Data Engineering**, v. 29, n. 12, p. 2724–2743, 2017.

WANG, Z. et al. Cross-lingual knowledge graph alignment via graph convolutional networks. **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018**, p. 349–357, 2018.

XU, H. et al. Text classification with topic-based word embedding and Convolutional Neural Networks. **ACM-BCB 2016 - 7th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics**, p. 88–97, 2016.

YU, H. et al. A relationship extraction method for domain knowledge graph construction. **World Wide Web**, v. 23, n. 2, p. 735–753, 2020.

ZHANG, H. The Optimality of Naive Bayes. **AA 1.2**, 2004.

ZHANG, H.; LI, D. Naïve Bayes Text Classifier. n. 3, p. 708–711, 2007.

ZHANG, Z.; HUANG, J.; TAN, Q. Association rules enhanced knowledge graph attention network. **arXiv**, 2020.

ZOU, H.; YANG, J.; WU, X. Unsupervised Energy-based Adversarial Domain Adaptation for Cross-domain Text Classification. p. 1208–1218, 2021.

APÊNDICE A – Cálculos do *ranking* final a partir dos métodos que constituem o modelo

indice	Ideia	Proj	O	P	V	NB_O	NB_P	NB_V	KG_O	KG_P	KG_V	WE_O	WE_P	WE_V	Res_NB	Res_KG	Res_WE	Rank	G(O)	G(P)	G(V)
1	1	0	1	0	0	1	0	0	1	1	0	0	0	0	1	1	0	41,67	22%	11%	0%
2	1	0	1	0	0	1	1	1	0	0	1	0	1	0	1	1	1	66,67	11%	22%	22%
3	1	0	0	1	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
4	1	0	0	1	0	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
5	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
6	1	0	1	0	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
7	1	0	0	1	0	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
8	1	0	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
9	1	0	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
10	1	0	1	1	0	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
11	1	0	0	1	0	1	1	1	0	0	0	0	1	1	1	0	1	58,33	11%	22%	22%
12	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	91,67	33%	22%	33%
13	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
14	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
15	1	0	0	1	1	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
16	1	0	1	0	0	0	0	0	1	1	1	1	1	0	0	1	1	58,33	22%	22%	11%
17	1	0	0	1	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
18	1	0	1	0	0	1	1	1	0	1	0	0	0	0	1	1	0	50,00	11%	22%	11%
19	1	0	1	0	0	1	1	1	0	0	0	1	1	1	1	0	1	66,67	22%	22%	22%
20	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	91,67	33%	22%	33%
21	1	0	1	0	0	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
22	1	0	1	1	0	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
23	1	0	1	0	0	1	1	1	1	0	0	0	1	0	1	1	1	66,67	22%	22%	11%
24	1	0	1	1	0	1	1	0	1	1	1	0	0	0	1	1	0	58,33	22%	22%	11%
25	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	25,00	11%	0%	11%
26	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	91,67	22%	33%	33%

27	1	0	1	0	0	1	1	1	1	0	1	0	1	0	1	1	1	75,00	22%	22%	22%
28	1	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
29	1	0	0	1	0	0	0	1	1	0	0	0	1	0	1	1	1	50,00	11%	11%	11%
30	1	0	1	0	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
31	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
32	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	91,67	33%	33%	22%
33	1	0	0	1	0	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
34	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
35	1	0	1	0	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
36	1	0	1	1	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
37	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
38	1	0	1	0	1	1	1	1	1	0	0	0	0	0	1	1	0	50,00	22%	11%	11%
39	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
40	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0	33,33	11%	11%	11%
41	1	0	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
42	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
43	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
44	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
45	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	91,67	33%	33%	22%
46	1	0	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
47	1	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	75,00	22%	22%	22%
48	1	0	0	1	1	1	1	1	1	1	1	0	0	0	1	1	0	66,67	22%	22%	22%
49	1	0	0	1	0	1	1	1	1	1	0	0	0	0	1	1	0	58,33	22%	22%	11%
50	1	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	83,33	22%	33%	22%
51	1	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	75,00	22%	22%	22%
52	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
53	1	0	0	1	0	1	1	1	1	0	1	0	0	0	1	1	0	58,33	22%	11%	22%
54	1	0	1	0	0	1	1	1	1	0	1	1	0	0	1	1	1	75,00	33%	11%	22%

55	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	100,00	33%	33%	33%
56	1	0	1	0	0	1	1	1	0	0	0	0	1	1	1	0	1	58,33	11%	22%	22%
57	0	1	0	0	0	0	0	0	0	1	0	1	1	1	0	1	1	50,00	11%	22%	11%
58	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	41,67	0%	22%	11%
59	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	16,67	11%	0%	0%
60	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
61	0	1	0	0	0	0	1	0	1	1	0	0	1	0	1	1	1	58,33	11%	33%	0%
62	0	1	0	0	0	0	0	0	1	0	1	1	1	1	0	1	1	58,33	22%	11%	22%
63	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	41,67	22%	11%	0%
64	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	66,67	22%	22%	22%
65	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	25,00	11%	11%	0%
66	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	58,33	22%	22%	11%
67	0	1	0	0	0	0	0	0	1	1	1	0	1	0	0	1	1	50,00	11%	22%	11%
68	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	16,67	11%	0%	0%
69	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	25,00	11%	0%	11%
70	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	16,67	11%	0%	0%
71	0	1	0	0	0	0	0	0	1	0	1	1	1	1	0	1	1	58,33	22%	11%	22%
72	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	25,00	11%	11%	0%
73	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	25,00	11%	11%	0%
74	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	33,33	0%	22%	0%
75	0	1	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	41,67	11%	11%	11%
76	0	1	0	0	0	0	0	0	1	0	1	1	1	0	0	1	1	50,00	22%	11%	11%
77	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	16,67	0%	11%	0%
78	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	16,67	11%	0%	0%
79	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
80	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	25,00	11%	0%	11%
81	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
82	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00	0%	0%	0%

83	0	1	0	0	0	0	0	0	1	0	1	1	1	1	0	1	1	58,33	22%	11%	22%
84	0	1	0	0	0	0	0	0	1	0	0	1	1	1	0	1	1	50,00	22%	11%	11%
85	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	58,33	22%	22%	11%
86	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00	0%	0%	0%
87	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
88	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	16,67	11%	0%	0%
89	0	1	0	0	0	0	0	0	1	0	1	1	1	0	0	1	1	50,00	22%	11%	11%
90	0	1	0	0	0	0	0	0	1	1	1	0	1	0	0	1	1	50,00	11%	22%	11%
91	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	1	41,67	11%	11%	11%
92	0	1	0	0	0	0	0	0	1	0	1	1	1	0	0	1	1	50,00	22%	11%	11%
93	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	33,33	11%	11%	0%
94	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
95	0	1	0	0	0	0	0	0	1	1	0	1	1	0	0	1	1	50,00	22%	22%	0%
96	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	1	1	41,67	11%	22%	0%
97	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
98	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	33,33	11%	11%	11%
99	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	25,00	11%	11%	0%
100	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00	0%	0%	0%
101	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	66,67	22%	22%	22%
102	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	16,67	0%	11%	0%
103	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	58,33	22%	22%	11%
104	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	1	41,67	11%	11%	11%
105	0	1	0	0	0	0	0	0	1	1	1	0	1	0	0	1	1	50,00	11%	22%	11%
106	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00	0%	0%	0%
107	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	25,00	0%	11%	11%
108	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	58,33	22%	22%	11%
109	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	58,33	22%	22%	11%
110	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	66,67	22%	22%	22%

111	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	16,67	11%	0%	0%
112	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	16,67	11%	0%	0%

APÊNDICE B – *Ranking* final ordenado

Posição no ranking	Valor do ranking	Índice	Ideia (1) / Projeto (0)	G(O)	G(P)	G(V)
1	100	5	1	33,33%	33,33%	33,33%
2	100	13	1	33,33%	33,33%	33,33%
3	100	31	1	33,33%	33,33%	33,33%
4	100	34	1	33,33%	33,33%	33,33%
5	100	42	1	33,33%	33,33%	33,33%
6	100	44	1	33,33%	33,33%	33,33%
7	100	52	1	33,33%	33,33%	33,33%
8	100	55	1	33,33%	33,33%	33,33%
9	91,66666667	12	1	33,33%	22,22%	33,33%
10	91,66666667	20	1	33,33%	22,22%	33,33%
11	91,66666667	26	1	22,22%	33,33%	33,33%
12	91,66666667	32	1	33,33%	33,33%	22,22%
13	91,66666667	45	1	33,33%	33,33%	22,22%
14	83,33333333	8	1	22,22%	33,33%	22,22%
15	83,33333333	9	1	22,22%	33,33%	22,22%
16	83,33333333	28	1	22,22%	33,33%	22,22%
17	83,33333333	41	1	22,22%	33,33%	22,22%
18	83,33333333	43	1	22,22%	33,33%	22,22%
19	83,33333333	46	1	22,22%	33,33%	22,22%
20	83,33333333	50	1	22,22%	33,33%	22,22%
21	75	27	1	22,22%	22,22%	22,22%
22	75	47	1	22,22%	22,22%	22,22%
23	75	51	1	22,22%	22,22%	22,22%
24	75	54	1	33,33%	11,11%	22,22%
25	66,66666667	2	1	11,11%	22,22%	22,22%
26	66,66666667	4	1	22,22%	22,22%	22,22%
27	66,66666667	7	1	22,22%	22,22%	22,22%
28	66,66666667	10	1	22,22%	22,22%	22,22%
29	66,66666667	14	1	22,22%	22,22%	22,22%
30	66,66666667	15	1	22,22%	22,22%	22,22%
31	66,66666667	19	1	22,22%	22,22%	22,22%
32	66,66666667	21	1	22,22%	22,22%	22,22%
33	66,66666667	22	1	22,22%	22,22%	22,22%
34	66,66666667	23	1	22,22%	22,22%	11,11%
35	66,66666667	33	1	22,22%	22,22%	22,22%
36	66,66666667	37	1	22,22%	22,22%	22,22%
37	66,66666667	39	1	22,22%	22,22%	22,22%
38	66,66666667	48	1	22,22%	22,22%	22,22%
39	66,66666667	64	0	22,22%	22,22%	22,22%
40	66,66666667	101	0	22,22%	22,22%	22,22%
41	66,66666667	110	0	22,22%	22,22%	22,22%
42	58,33333333	3	1	22,22%	11,11%	22,22%

43	58,33333333	6	1	22,22%	11,11%	22,22%
44	58,33333333	11	1	11,11%	22,22%	22,22%
45	58,33333333	16	1	22,22%	22,22%	11,11%
46	58,33333333	17	1	22,22%	11,11%	22,22%
47	58,33333333	24	1	22,22%	22,22%	11,11%
48	58,33333333	30	1	22,22%	11,11%	22,22%
49	58,33333333	35	1	22,22%	11,11%	22,22%
50	58,33333333	36	1	22,22%	11,11%	22,22%
51	58,33333333	49	1	22,22%	22,22%	11,11%
52	58,33333333	53	1	22,22%	11,11%	22,22%
53	58,33333333	56	1	11,11%	22,22%	22,22%
54	58,33333333	61	0	11,11%	33,33%	0,00%
55	58,33333333	62	0	22,22%	11,11%	22,22%
56	58,33333333	66	0	22,22%	22,22%	11,11%
57	58,33333333	71	0	22,22%	11,11%	22,22%
58	58,33333333	83	0	22,22%	11,11%	22,22%
59	58,33333333	85	0	22,22%	22,22%	11,11%
60	58,33333333	103	0	22,22%	22,22%	11,11%
61	58,33333333	108	0	22,22%	22,22%	11,11%
62	58,33333333	109	0	22,22%	22,22%	11,11%
63	50	18	1	11,11%	22,22%	11,11%
64	50	29	1	11,11%	11,11%	11,11%
65	50	38	1	22,22%	11,11%	11,11%
66	50	57	0	11,11%	22,22%	11,11%
67	50	67	0	11,11%	22,22%	11,11%
68	50	76	0	22,22%	11,11%	11,11%
69	50	84	0	22,22%	11,11%	11,11%
70	50	89	0	22,22%	11,11%	11,11%
71	50	90	0	11,11%	22,22%	11,11%
72	50	92	0	22,22%	11,11%	11,11%
73	50	95	0	22,22%	22,22%	0,00%
74	50	105	0	11,11%	22,22%	11,11%
75	41,66666667	1	1	22,22%	11,11%	0,00%
76	41,66666667	58	0	0,00%	22,22%	11,11%
77	41,66666667	63	0	22,22%	11,11%	0,00%
78	41,66666667	75	0	11,11%	11,11%	11,11%
79	41,66666667	91	0	11,11%	11,11%	11,11%
80	41,66666667	96	0	11,11%	22,22%	0,00%
81	41,66666667	104	0	11,11%	11,11%	11,11%
82	33,33333333	40	1	11,11%	11,11%	11,11%
83	33,33333333	60	0	11,11%	11,11%	11,11%
84	33,33333333	74	0	0,00%	22,22%	0,00%
85	33,33333333	79	0	11,11%	11,11%	11,11%
86	33,33333333	81	0	11,11%	11,11%	11,11%
87	33,33333333	87	0	11,11%	11,11%	11,11%

88	33,33333333	93	0	11,11%	11,11%	0,00%
89	33,33333333	94	0	11,11%	11,11%	11,11%
90	33,33333333	97	0	11,11%	11,11%	11,11%
91	33,33333333	98	0	11,11%	11,11%	11,11%
92	25	25	1	11,11%	0,00%	11,11%
93	25	65	0	11,11%	11,11%	0,00%
94	25	69	0	11,11%	0,00%	11,11%
95	25	72	0	11,11%	11,11%	0,00%
96	25	73	0	11,11%	11,11%	0,00%
97	25	80	0	11,11%	0,00%	11,11%
98	25	99	0	11,11%	11,11%	0,00%
99	25	107	0	0,00%	11,11%	11,11%
100	16,66666667	59	0	11,11%	0,00%	0,00%
101	16,66666667	68	0	11,11%	0,00%	0,00%
102	16,66666667	70	0	11,11%	0,00%	0,00%
103	16,66666667	77	0	0,00%	11,11%	0,00%
104	16,66666667	78	0	11,11%	0,00%	0,00%
105	16,66666667	88	0	11,11%	0,00%	0,00%
106	16,66666667	102	0	0,00%	11,11%	0,00%
107	16,66666667	111	0	11,11%	0,00%	0,00%
108	16,66666667	112	0	11,11%	0,00%	0,00%
109	0	82	0	0,00%	0,00%	0,00%
110	0	86	0	0,00%	0,00%	0,00%
111	0	100	0	0,00%	0,00%	0,00%
112	0	106	0	0,00%	0,00%	0,00%