Wellington Eufrasio Camargo

A Construção de Solução de Visualização da Informação para Prestação de Contas e Exploração da Execução Orçamentária

Bauru, São Paulo, Brasil

Junho 2023

Wellington Eufrasio Camargo

A Construção de Solução de Visualização da Informação para Prestação de Contas e Exploração da Execução Orçamentária

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação – PPGCC, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.
Área de concentração: Computação Aplicada.
Linha de pesquisa: Sistemas de Informação.

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. José Remo Ferreira Brega

Bauru, São Paulo, Brasil

Junho 2023

C172c

Camargo, Wellington Eufrasio
A Construção de Solução de Visualização da Informação para Prestação de Contas e Exploração da Execução Orçamentária / Wellington Eufrasio Camargo. -- Bauru, 2023
161 p. : il., tabs., fotos, mapas

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru
Orientador: José Remo Ferreira Brega

1. Visualização da informação. 2. Sistemas de informação gerencial. 3. Transparência na administração pública. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

Faculdade de Ciências

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE WELLINGTON EUFRASIO CAMARGO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 05 dias do mês de maio do ano de 2023, às 15:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de WELLINGTON EUFRASIO CAMARGO, intitulada **A Construção de Solução de Visualização da Informação para Prestação de Contas e Exploração da Execução Orçamentária**. A Comissão Examinadora foi constituida pelos seguintes membros: Prof. Dr. JOSE REMO FERREIRA BREGA (Orientador(a) - Participação Virtual) do(a) Departamento de Computacao / UNESPCampus de Bauru, Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Participacão Virtual) do(a) FC / UNESPBauru SP, Prof. Dr. ILDEBERTO APARECIDO RODELLO (Participação Virtual) do(a) Departamento de Administração da FEA / Universidade de São Paulo. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: APROVADO. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. JOSE REMO FERREIRA BREGA

Para Olívia...

"Nunca estou realmente satisfeita por entender alguma coisa; porque, entenda-o tão bem quanto eu possa, minha compreensão só pode ser uma fração infinitesimal de tudo que eu quero entender sobre as muitas conexões e relações que me ocorrem, como o assunto em questão foi pensado ou chegado pela primeira vez, etc., etc."

Ada Lovelace

Agradecimentos

Primeiro, agradeço à Deus, por me dar a vida, a saúde e a capacidade, sem as quais não seria possível realizar este trabalho. Agradeço especialmente minha esposa Thamires e nossa filha Olívia, por todo amor da vida e por todo suporte familiar oferecido neste jornada acadêmica. Agradeço muito aos companheiros de trabalho, que participaram e apoiaram a realização deste estudo, Luttgardes, Willians, Ariane, Camila, Manoel, Rubens e José Henrique. Agradeço ao meu orientador, Prof. Dr. José Remo Ferreira Brega, por aceitar o desafio de percorrermos esta jornada acadêmica juntos. Igualmente expresso meus agradecimentos aos docentes do Programa de Pós-Graduação em Ciência da Computação e à secretaria do programa, que de alguma forma colaboraram com minha formação. E agradeço à você, por prestigiar este trabalho.

Resumo

Motivado por um problema real de mineração manual de texto realizada em registros financeiros e contábeis, conduzida para obter-se informações à serem utilizadas na construção de gráficos estáticos em relatórios gerenciais, este trabalho apresenta um estudo de caso sobre a construção e integração de solução de visualização da informação, com objetivo de elucidar a execução orçamentária de uma universidade pública do Estado de São Paulo, transformando a complexa e burocrática prestação de contas em informações visuais e de fácil entendimento. Partindo-se da questão de como a Informática pode contribuir na elucidação de registros públicos contábeis genéricos e pouco informativos, realizou-se o pré-processamento de dados dos registros financeiros e contábeis da universidade, utilizando-se a classificação automática de texto. Desta forma, foi possível atribuir uma finalidade qualitativa ao emprego de recursos públicos, demonstrando-se para que o recurso financeiro está sendo utilizado, atribuindo-se também uma categoria discriminativa, que possibilita eliminar generalizações e estabelecer efetivamente em que o recurso está sendo empregado. A realização deste estudo de caso confirmou a hipótese de que ferramentas de visualização da informação, criadas com informações padronizadas por ferramenta de classificação, são eficazes para transparência pública e são eficientes para elucidação da execução orçamentária. A avaliação da solução de implantada em uma das faculdades da universidade, demonstrou que a solução desenvolvida atende tanto as necessidades de informações gerenciais para os gestores de centros de custo, quanto a possibilidade do acompanhamento da execução orçamentária na universidade. O desenvolvimento do trabalho demonstrou que conceitos, técnicas e métodos relacionados com aprendizado de máquina, integram-se ao processo de construção de ferramentas de visualização da informação, contribuindo especialmente no pré-processamento dos dados a serem utilizados nas ferramentas de visualização. Os resultados obtidos levam a conclusão que é possível estabelecer um processo para a construção e integração de ferramentas de visualização em um sistema institucional legado e, que a solução de visualização da informação desenvolvida é uma solução amigável e eficiente de análise visual para elucidação da prestação de contas e exploração da execução orçamentária na universidade.

Palavras-chave: Visualização da Informação; Classificação de Texto; Execução Orçamentária; Prestação de Contas; Transparência Pública.

Abstract

Motivated by a real problem of manual text mining carried out in financial and accounting records, performed to obtain information to be used in the construction of static graphs in management reports, this work presents a case study on the construction and integration of an visualization of information solution, with the objective of elucidating the budget execution of a public university in the State of São Paulo, transforming the complex and bureaucratic rendering of accounts into easy to understand visual information. Starting from the question of how Computing can contribute to the elucidation of generic and uninformative accounting public records, the data from the university's financial and accounting records was preprocessed using the automatic text classification. In this way, it was possible to attribute a qualitative purpose to the use of public resources, demonstrating in what the financial resource is used, also assigning a discriminative category, that makes it possible to eliminate generalizations and effectively establish for what the resource is used. The realization of this case study confirmed that information visualization tools created with standardized information by classification tool, are effective for public transparency and efficient for elucidation of budget execution. The evaluation of the solution implemented in one of the university's college, demonstrated that the developed solution meets both the needs of management information for cost center managers, as well as the possibility of monitoring the budget execution at the university. The development of the work demonstrated that concepts, techniques and methods related to machine learning could be integrated into the process of building information visualization tools, contributing especially to the preprocessing of data to be used in the visualization tools. The results obtained lead to the conclusion that it is possible to establish a process for the construction and integration of visualization tools in a legacy system and that the visualization of information solution developed is a friendly and efficient visual analytics solution for accountability and exploration of budget execution in the university.

Keywords: Visualization of Information; Text Classification; Budget Execution; Accountability; Public Transparency.

Lista de ilustrações

Lista de tabelas

Lista de algoritmos

Lista de abreviaturas e siglas

ACM	*Association for Computing Machinery*
ACM CSS	*ACM Computing Classification System*
AI	Inteligência Artificial
BI	*Business Intelligence*
BoW	*Bag of Words*
BR	*Binary Relevance*
CART	*Classification And Regression Tree*
CNN	*Convolutional Neural Network*
CSS	*Cascading Style Sheets*
D3	*D3.js – Data-Driven Documents*
DL	*Deep Learning*
DM	*Data Mining*
DT	*Decision Tree*
ES	*Expert System*
FEB	Faculdade de Engenharia de Bauru
FN	Falso Negativo
FNN	*Fuzzy Neural Network*
FP	Falso Positivo
GA	*Genetic Algorithm*
GMDH	*Group Method of Data Handling*
GP	*Gaussian Process*
GranDSI-BR	Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil
GRU	*Gated Recurrent Units*
Harvard GI	*Harvard General Inquirer*
HTML	*Hypertext Markup Language*

IA	*Intelligent Agents*
ICMS	Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação
IEEE	*The Institute of Electrical and Electronics Engineers*
IET	*The Institution of Engineering and Technology*
IHC	Interação Humano-Computador
IR	*Information Retrieval*
JEL	*Journal of Economic Literature*
k-NN	*k-Nearest Neighbour*
KWIC	*Keyword in Context*
LDA	*Latent Dirichlet Allocation*
LM dictionary	*Loughran and Mcdonald dictionary*
LP	*Label Powerset*
LR	*Logistic Regression*
LSTM	*Long Short-Term Memory*
MDL	*Minimum Description Length*
ML-CKNN	*Multilabel Categorical K-Nearest Neighbor*
ML	*Machine Learning*
MLN	*Mandatory Leaf Node Prediction*
NB	*Naive Bayes*
NER	*Named Entity Recognition*
NLP	*Natural Language Processing*
NMLN	*Non-mandatory Leaf Node Prediction*
NNS	*Nearest Neighbour Separation*
PA	*Passive Aggressive*
PCA	*Principal Component Analysis*
PoS	*Part-of-Speech*
Q&A	*Question & Answering (Q&A)*

RE	*Relation Extraction*
RNN	*Recurrent Neural Network*
RSL	Revisão Sistemática da Literatura
SEC	*United States Securities and Exchange Commission's*
SF	*Slot Filling*
SGD	*Stochastic Gradient Descent*
SOM	*Self-Organizing Maps*
SRL	*Semantic Role Labeling*
SVD	*Singular Value Decomposition*
SVM	*Support Vector Machine*
TF–IDF	*Term Frequency – Inverse Document Frequency*
TI	Tecnologia da Informação
TM	*Text Mining*
UNA-SUS	Universidade Aberta do SUS
UNESP	Universidade Estadual Paulista "Júlio de Mesquita Filho"
UNICAMP	Universidade Estadual de Campinas
USP	Universidade de São Paulo
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XMTC	*Extreme Multi-label Text Classification*

Sumário

1 Introdução

A análise de registros financeiros e contábeis de uma universidade pública paulista foi o problema motivador deste trabalho. Nestes registros, encontram-se informações de diversas origens, que podem ser categorizadas de forma distinta e mais amigável do que os códigos contábeis disponíveis na legislação vigente. Além disso, é vasta a quantidade de registros que não possuem informação contábil relacionada à *"Classificação da Despesa Orçamentária por Natureza"* estabelecida pela Secretaria da Fazenda e Planejamento (2023), ou possuem classificações que estão vinculadas a elementos genéricos e pouco informativos como: *"Outros serviços..."*, *"Outros materiais..."*, *"Outras taxas..."*, etc.

De forma semelhante a categorização relacionada com a natureza das despesas, destaca-se ainda nos registros analisados, a presença de classificação funcional programática. Classificação que atribui aos registros, informações relacionadas à finalidade de destino do recurso financeiro empregado, que contabilmente são descritas por finalidades estabelecidas em lei. Contudo, é frequente a utilização de classificação relacionada ao objetivo geral da instituição, por exemplo: *"Ensino de graduação nas universidades públicas"* (SÃO PAULO, 2020). Observa-se que esta forma de classificação generalista, pode não refletir a real finalidade do emprego dos recursos financeiros na universidade.

Apesar do exposto, é frequente a presença de informações adicionais relacionadas aos registros contábeis, que indicam a possibilidade de categorizações distintas ou mais específicas do que as dispostas na legislação relacionada. Estas informações remetem aos procedimentos que deram origem ao lançamento contábil, contendo dados que justifiquem a necessidade de determinadas despesas, dados oriundos da pessoa solicitante, que alimenta os sistemas institucionais com justificativas redigidas em linguagem natural. Desta forma, considera-se a possibilidade de classificar os registros financeiros e contábeis, categorizando-os de forma mais simples do que a legislação vigente. Considerando também, o estabelecimento automático de vínculo de cada registro contábil, com as atividades-fim que formam o tripé base de qualquer universidade (Ensino, Pesquisa e Extensão), ou com atividades-meio como Infraestrutura e Administração.

Dentro da grande área do conhecimento que estuda as relações da Interação Humano-Computador (IHC), a Visualização da Informação é a área que dedicada-se ao estudo de sistemas de visualização baseados em computador. Estudo que contempla, não só, mas ao menos, o design, o processo, a construção e a avaliação de ferramentas de visualização (MUNZNER, 2014).

Esta monografia é a apresentação do estudo de caso, da criação e integração de ferramentas de visualização da informação, ao sistema legado de gestão administrativa de uma universidade pública do Estado de São Paulo. Sendo estas ferramentas de visualização, baseadas em informações padronizadas por algoritmo de classificação de texto, desenvolvido aplicando-se abordagens, conceitos, técnicas e métodos da área de Mineração de Texto.

1.1 Objetivos

O objetivo geral do trabalho é a elucidação de registros públicos financeiros e contábeis, transformando em informações visuais a classificação textual legalmente utilizada. Informações visuais que reflitam tanto categorias como finalidades, mais específicas e amigáveis sob a perspetiva humana, no intuito de qualificar e discriminar a complexa e burocrática execução orçamentária da instituição. Transmitindo de forma simples as informações de como e com qual finalidade, estão sendo empregados os recursos financeiros da universidade.

Neste sentido, o objetivo específico deste trabalho é o desenvolvimento de solução que atenda tanto as necessidades de informações gerenciais para o gestor do centro de custo, quanto a possibilidade de acompanhamento da execução orçamentária na universidade. Solução desenvolvida utilizando ferramentas de visualização da informação, que permitam os usuários explorarem registros contábeis da universidade. No intuito de facilitar o entendimento dos recursos financeiros recebidos, de como estes recursos foram utilizados e, com qual finalidade os recursos foram empregados.

1.2 Relevância da Pesquisa

A iniciativa da Sociedade Brasileira de Computação (SBC), em prospectar os principais desafios de pesquisa na área da Ciência da Computação, evidenciou-se com os seminários *"Grandes Desafios em Pesquisa em Computação no Brasil"*. Esta iniciativa da SBC gerou impacto positivo na comunidade científica, permitindo o desenvolvimento concreto de ações que abordassem os temas propostos nos seminários (BOSCARIOLI; ARAUJO; MACIEL, 2017).

Inspirada nesta iniciativa e considerando o amadurecimento da comunidade científica de Sistemas de Informação, a Comissão Especial de Sistemas de Informação (CE-SI) da SBC, lançou em 2016 o seminário *"Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil (GranDSI-BR)"* . Cujo principal objetivo é identificar os desafios a serem enfrentados nas pesquisas da área para a próxima década. Boscarioli, Araujo e Maciel (2017) apontam nesta primeira edição do seminário, quatro grandes desafios de pesquisa em sistemas de informação para o decênio 2016-2026:

- Desafio 1: Sistemas de Sistemas de Informação.

  > No mundo aberto, globalizado e conectado, os sistemas de informação não apenas suportam uma grande diversidade de domínios de aplicação, como negócios, saúde e resposta a crises, mas executam várias tarefas e funcionalidades complexas. Os Sistemas de Sistemas de Informação (SoIS - *Systems of Information Systems*) são um tipo específico de Sistemas de Sistemas (SoS - *Systems of Systems*) que apresenta novos desafios para o desenvolvimento de Sistemas de Informação (IS - *Information Systems*) e a comunidade de pesquisa. Os SoIS exibem todas as características de SoS, com uma forte característica adicional relacionada com a natureza do negócio. SoIS são compostos de vários IS que combinam suas capacidades (BOSCARIOLI; ARAUJO; MACIEL, 2017, adaptado, tradução livre).

- Desafio 2: Sistemas de Informação e os Desafios do Mundo Aberto.

  > O mundo é uma rede. O desafio é compreender sua dinâmica e propor, construir e compreender o impacto dos sistemas de informação para

> apoiá-la. Uma longa lista de aspectos deve ser considerada ao associar sistemas de informação ao mundo aberto e virtual. Isso inclui: mobilidade, colaboração, capacitação, interoperabilidade, compartilhamento de conhecimento, escalabilidade, transparência, privacidade, segurança, flexibilidade, valor, confiabilidade, diversidade, licenciamento... a lista é infinita. As novas tendências tecnológicas também devem ser levadas em consideração: dados abertos e vinculados, redes sociais, sistemas multi-agentes, apenas para citar alguns. **O mundo aberto é verdadeiro e necessário para diferentes domínios de aplicação, desde a prestação de serviços à inovação, incluindo o acesso da sociedade à informação e a participação, tanto no setor público quanto no privado**. Relacionamentos diferentes entre consumidores e fornecedores estão surgindo. Qualquer um pode ser um produtor, qualquer um pode ser um consumidor no mundo aberto. **Novos ecossistemas surgem deste mundo conectado e novas abordagens para projetar e fornecer sistemas de informação para apoiar esses ecossistemas são necessários, desafiando a legislação brasileira, o governo, a indústria e os processos de produção do mercado e, o comportamento, a educação e cultura das pessoas** (BOSCARIOLI; ARAUJO; MACIEL, 2017, grifo nosso, tradução livre).

- Desafio 3: Complexidade dos sistemas de informação.

> Os Sistemas de Informação atuais e futuros compreendem vários componentes. Esses componentes podem ser outros sistemas, software ou sensores hospedados em diferentes plataformas computacionais. Devido à diversidade e quantidade de componentes, os IS estão se tornando cada vez mais complexos. No contexto dos sistemas de informação, a troca de informações e a interação entre os usuários frequentemente ocorrem em ambientes heterogêneos. A interoperabilidade é um requisito fundamental para apoiar as atividades em ambientes heterogêneos de forma eficiente e eficaz. Além disso, no que diz respeito à infraestrutura de tecnologia da informação para sistemas de informação, o suporte virtual e as plataformas de desenvolvimento estão mudando a forma como os clientes interagem com os dados e aplicativos (BOSCARIOLI; ARAUJO; MACIEL, 2017, adaptado, tradução livre).

- Desafio 4: Visão Sociotécnica de Sistemas de Informação.

> **Os sistemas de informação não são apenas software ou pessoas que usam software. Eles são a integração total de pessoas e tecnologia e a multiplicidade de relacionamentos que surgem dessa integração**. Os sistemas de informação hoje e nos próximos anos não podem ser projetados, desenvolvidos, pesquisados, usados ou aprendidos sem abordagens consistentes para lidar com a complexidade do sistema sociotécnico que nossa sociedade é e continuará a ser. Resolver efetivamente os problemas dos sistemas de informação significa desenvolver competências em pesquisa, educação e na comunidade profissional de SI para **compreender plenamente o que é uma visão sociotécnica e aplicar de forma consistente métodos e práticas interdisciplinares para entender e resolver problemas do mundo real** (BOSCARIOLI; ARAUJO; MACIEL, 2017, grifo nosso, tradução livre).

O objeto de estudo apresentado neste trabalho são dados financeiros e contábeis de uma universidade pública do Estado de São Paulo. Dados que representam a distribuição e a execução orçamentária da universidade, consolidando informações hierarquicamente desde o centro de custo (micro-gestão) até a universidade (macro-gestão). Desta forma, identifica-se o enquadramento deste trabalho nos desafios 2 e 4 propostos por Boscarioli, Araujo e Maciel (2017). Considerando os grandes desafios de pesquisa em sistemas de informação no Brasil, considerando também a complexidade burocrática-legislativa da gestão orçamentária da instituição observada,

considerando ainda, a vasta generalização de informações contidas nos registros contábeis, este estudo demonstra-se relevante por abordar problemas que são desafios contemporâneos na área de Sistemas de Informação.

1.3 Abordagens Metodológicas

Diferentes abordagens metodológicas foram utilizadas de acordo com as fases de desenvolvimento deste estudo. Estas abordagens serão descritas nas próximas seções.

1.3.1 Visão Geral sobre a Metodologia do Trabalho

Este trabalho é a materialização de uma pesquisa qualitativa, realizada com o método de estudo de caso, para aprofundamento nos temas envolvidos na implantação de ferramentas e técnicas de visualização da informação integradas à sistemas institucionais de uma universidade.

O método de estudo de caso demonstra-se adequado para profunda investigação de como e porque ocorrem os eventos relacionados com o objeto de pesquisa. É uma investigação empírica, que permite o estudo de um fenômeno contemporâneo dentro de seu contexto da vida real. Dentre as finalidades da realização de um estudo de caso, destacam-se para esta pesquisa os propósitos de formular hipóteses, desenvolver teorias e descrever a situação do contexto em que está sendo realizada a investigação (GIL, 2002; YIN, 2015).

1.3.2 Metodologia para o Referencial Teórico

O referencial teórico desta pesquisa foi estabelecido pelo método da Revisão Sistemática da Literatura (RSL) e complementado por pesquisas bibliográficas necessárias para abordagem dos temas identificados.

Kitchenham e Charters (2007) definem a RSL como uma forma de avaliar e interpretar todas as pesquisas relevantes, que estão disponíveis para uma questão de pesquisa específica, uma área de tópico ou um fenômeno de interesse. As revisões sistemáticas visam apresentar uma avaliação justa de um tópico de pesquisa, através de uma metodologia de pesquisa rigorosa, confiável e auditável.

A RSL foi realizada com objetivo de identificar contribuições acadêmicas e científicas no estudo das relações entre aprendizado de máquina e visualização da informação, ao se trabalhar com conteúdo textual relacionado com informações financeiras e contábeis. Os Apêndices B e C apresentam a revisão sistemática realizada, discorrendo detalhadamente sobre seu planejamento, condução e resultados.

1.3.3 Metodologia para Avaliação do Método de Classificação

Realizou-se um estudo empírico para avaliar a efetividade do método de classificação proposto neste trabalho. O estudo realizado contemplou dois tipos de problemas de classificação e comparou o desempenho do método proposto com seis algoritmos de referência na tarefa de classificação de texto.

Um conjunto de dados experimental foi criado para ser utilizado nos testes de classificação. Todos os registros obtidos foram manualmente classificados por especialistas do domínio da aplicação, realizando-se a atribuição de uma categoria e uma finalidade à cada registro. O *dataset* estabelecido foi utilizado para os testes com a implementação do classificador apresentado no Capítulo 5, realizando-se comparações com implementações de algoritmos de classificação do tipo aprendizado de máquina disponíveis na ferramenta Weka Workbench (EIBE; HALL; WITTEN, 2016).

O Capítulo 6 apresenta descrição detalhada dos experimentos realizados, discorrendo sobre os dados utilizados, descrevendo parâmetros definidos e apresentando os resultados obtidos.

1.3.4 Metodologia para Avaliação da Solução de Visualização da Informação

Conforme estabelecido por Kirk (2012), detalhado no Item 3.2.6 do referencial teórico, a avaliação de ferramentas de visualização tem como objetivo identificar a eficácia e o impacto das visualizações criadas. O autor destaca destaca algumas questões consideradas como principais tópicos de interesse nesta avaliação (e.g., Houve reação positiva à visualização? Usuários foram capazes de consumir ou descobrir ideias de forma eficaz? Quais os problemas que as pessoas experimentam, se houveram?). Destacando também o *feedback* do cliente e o *feedback* não estruturado, como meios de se obter respostas para estes tópicos de avaliação.

Desta forma, a entrevista aberta foi o método de avaliação escolhido para obter-se respostas qualitativas à solução apresentada. Foram realizadas sete sessões online[1] para apresentação da solução e treinamento dos usuários, sendo duas sessões de apresentação da solução para o público-alvo e cinco sessões de treinamento específicas para usuários gestores de centros de custo.

Durante a realização das sessões, participantes ofereceram *feedbacks* não estruturados em suas manifestações verbais. Após a apresentação do conteúdo dedicado à sessão, conduziu-se a entrevista aberta, com o orador passando a palavra aos demais participantes para que estes se manifestassem sobre a solução apresentada.

As repostas e manifestações dos participantes foram registradas nas gravações, sendo algumas transcritas neste trabalho como citações de membros devidamente qualificados nos detalhes das sessões realizadas.

◆ Sessão de Apresentação I — Sem Gravação — Realizada em 01 de julho de 2021.

Apresentação preliminar conduzida pelo autor deste trabalho, qualificado como o assistente de informática membro do grupo de trabalho envolvido neste estudo de caso. A apresentação teve como público-alvo outros quatro integrantes do grupo de trabalho, tendo o diretor administrativo juntamente com seu assessor representando a área de gestão da faculdade, enquanto o diretor e o analista de informática representaram a área de tecnologia da informação.

Esta reunião não foi gravada, pois foi realizada como apresentação preliminar do trabalho, com objetivo de validar o protótipo da solução implementada.

[1] As sessões foram realizadas por meio da ferramenta Google Meet e foram gravadas utilizando-se da opção de gravação disponível na ferramenta.

O participante diretor da área administrativa ofereceu oralmente *feedback* de cliente.

◆ Sessão de Apresentação II — Gravada — Realizada em 14 de setembro de 2021.

Conduzida pelo grupo de trabalho envolvido neste estudo de caso, teve como público-alvo 24 pessoas responsáveis pela gestão de recursos financeiros da faculdade. Público composto por chefes de departamento, coordenadores de programas de pós-graduação, diretores de área e supervisores de seções.

Cinco participantes ofereceram *feedback* de forma oral e foram registradas mais nove interações por mensagens escritas no *chat*, das quais quatro eram dúvida ou *feedback*.

◆ Sessão de Treinamento I — Sem Gravação — Realizada em 23 de novembro de 2021.

Esperava-se a participação de três pessoas nesta sessão de treinamento. Contudo, apenas uma pessoa esteve presente, um assistente de suporte acadêmico que exercia a função de auxiliar de laboratório didático.

Esta sessão de treinamento não foi gravada, devido a manifestação contrária do participante. O participante informou não ser necessário realizar a gravação do treinamento, pelo fato de não ser o gestor responsável pelo departamento a qual o laboratório está vinculado e devido a participação do respectivo gestor em uma das próximas sessões de treinamento.

Não foram coletados *feedbacks* nesta sessão.

◆ Sessão de Treinamento II — Gravada — Realizada em 23 de novembro de 2021.

Reunião realizada com chefe de departamento, identificado neste trabalho pelo pseudônimo CDASPP. Participante manifestou *feedbacks* voluntários, expressando opinião sobre a usabilidade das ferramentas.

◆ Sessão de Treinamento III — Gravada — Realizada em 24 de novembro de 2021.

Os principais interessados da área de gestão administrativa da faculdade foram o público-alvo da terceira sessão de treinamento. Grupo de cinco pessoas composto por diretor de unidade universitária acompanhado de seu assessor; diretor da área administrativa também acompanhado por seu assessor; e, supervisor da seção de finanças.

Os participantes qualificados como diretor de unidade universitária e diretor administrativo, ofereceram oralmente *feedbacks* espontâneos sobre a solução apresentada. Também foram registradas três interações por mensagens via *chat*, mas nenhuma relacionada com dúvida ou *feedback*.

◆ Sessão de Treinamento IV — Gravada — Realizada em 24 de novembro de 2021.

Reunião realizada com supervisor de seção da área acadêmica, identificado pelo pseudônimo SSAAAAP. Participante manifestou *feedbacks* voluntários, expressando dúvidas e oferecendo exemplo sobre diferentes formas de classificação de finalidade para um mesmo item de compra.

◆ Sessão de Treinamento V — Gravada — Realizada em 25 de novembro de 2021.

Duas pessoas participaram da última sessão de treinamento, chefe de departamento, identificado neste trabalho pelo pseudônimo CDSDRO, e supervisor de seção da área acadêmica, identificado pelo pseudônimo SSAAEFL.

O participante qualificado como chefe de departamento manifestou *feedbacks* voluntários, expressando dúvidas, discorrendo sobre a funcionalidade da solução e comentando sobre a aderência da solução às necessidades de gestão. Também foi registrada uma participação de SSAAEFL, comentando sobre a origem das categorias e finalidades utilizadas para classificação de registros.

1.4 Organização do Trabalho

Com relação a forma, esta dissertação está organizada em nove capítulos e cinco apêndices. Os capítulos consistem no conteúdo principal da dissertação, enquanto os apêndices apresentam conteúdo suplementar gerado durante o desenvolvimento do trabalho.

O Capítulo 1 dedica-se à introdução geral, apresentando o tema da dissertação, seus objetivos e a relevância desta pesquisa. Em seguida, o Capítulo 2 discorre sobre o estudo de caso realizado, apresentando a instituição objeto de estudo, as hipóteses levantadas e descrevendo resumidamente as etapas e ações relacionadas em cada fase do estudo de caso. O Capítulo 3 apresenta conceitos fundamentais relacionados ao tema de pesquisa, necessários para melhor compreensão deste trabalho. O Capítulo 4 descreve os dois problemas de classificação de texto relacionados com este trabalho. Os Capítulos 5 e 6 dedicam-se à apresentação detalhada respectivamente da solução para classificação de texto e a avaliação da solução de classificação proposta. O Capítulo 7 apresenta a solução proposta para o problema tema desta pesquisa, discorrendo também sobre a implantação e avaliação da solução. Por fim, o Capítulo 8 realiza considerações finais e o Capítulo 9 conclui o conteúdo principal desta dissertação.

Com relação ao conteúdo, os próximos capítulos poderão apresentar termos em língua inglesa, sempre destacados de forma *itálica* e acompanhados de tradução livre. A escolha por esta forma de redação se justifica devido a internacionalização de conceitos, sendo estes amplamente conhecidos em inglês e/ou popularizados por siglas que refletem a escrita em língua inglesa.

2 Estudo de Caso

Este capítulo descreve brevemente o estudo de caso realizado, contextualizando o estudo e introduzindo etapas do trabalho, desde a concepção até a fase de execução.

2.1 Instituição Observada

A Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) é a instituição objeto de estudo deste trabalho. Uma grande universidade, composta por 34 unidades que estão em 24 cidades do Estado de São Paulo (APE, 2021a).

> A Unesp (Universidade Estadual Paulista) é uma das maiores e mais importantes universidades brasileiras, com destacada atuação no ensino, na pesquisa e na extensão de serviços à comunidade.
>
> Mantida pelo governo do estado de São Paulo, é uma das três universidades públicas de ensino gratuito, ao lado da USP (Universidade de São Paulo) e da Unicamp (Universidade Estadual de Campinas).
>
> Criada em 1976, a partir da reunião de institutos isolados de ensino superior que existiam em várias regiões do estado de São Paulo, a Unesp tem 34 unidades em 24 cidades, sendo 22 no interior; uma na capital; e uma no litoral paulista, mais especificamente na cidade de São Vicente.
>
> **Missão** Exercer sua função social por meio do ensino, da pesquisa e da extensão universitária, com espírito crítico e livre, orientados por princípios éticos e humanísticos. Promover a formação profissional compromissada com a qualidade de vida, a inovação tecnológica, a sociedade sustentável, a equidade social, os direitos humanos e a participação democrática. Gerar, difundir e fomentar o conhecimento, contribuindo para a superação de desigualdades e para o exercício pleno da cidadania.
>
> **Visão de futuro** Ser referência nacional e internacional de universidade pública multicâmpus, de excelência no ensino, na pesquisa e na extensão universitária, que forme profissionais e pesquisadores capazes de promover a democracia, a cidadania, os direitos humanos, a justiça social e a ética ambiental e que contribua para o letramento científico da sociedade e para a utilização pública da ciência. (UNESP, 2023).

A Figura 1 apresenta as cidades que possuem campus da Unesp.

Figura 1 – Distribuição geográfica dos campi da universidade no Estado de São Paulo.

Fonte: Extraído de UNESP (2023).

Em conjunto com as demais universidades mantidas pelo Governo do Estado de São Paulo, a Unesp possui dotação orçamentária própria na Lei Orçamentária Anual do Estado, legislação que estabelece valores e diretrizes para aplicação dos recursos financeiros da universidade (SÃO PAULO, 2020, p 499-501).

Os recursos financeiros para execução do orçamento planejado, são oriundos majoritariamente de quota-parte do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS), em conjunto com recursos obtidos de convênios com outros órgãos e instituições, complementados também por recursos próprios gerados na universidade (RANIERI, 2018; APE, 2021a).

A partir da dotação orçamentária atribuída pelo Estado e da previsão de arrecadação do ICMS no ano de exercício, a Unesp estabelece seu próprio orçamento anual, planejado hierarquicamente no sentido *top-down* (de cima para baixo, ou seja, da administração superior para as unidades administrativas), realizando a distribuição de dotações orçamentárias para cada uma de suas 34 unidades universitárias (APE, 2021b).

Constata-se então que a Unesp possui um único orçamento centralizado, mas com sua execução orçamentária distribuída em suas sedes administrativas, denominadas neste trabalho como unidades universitárias ou simplesmente unidades.

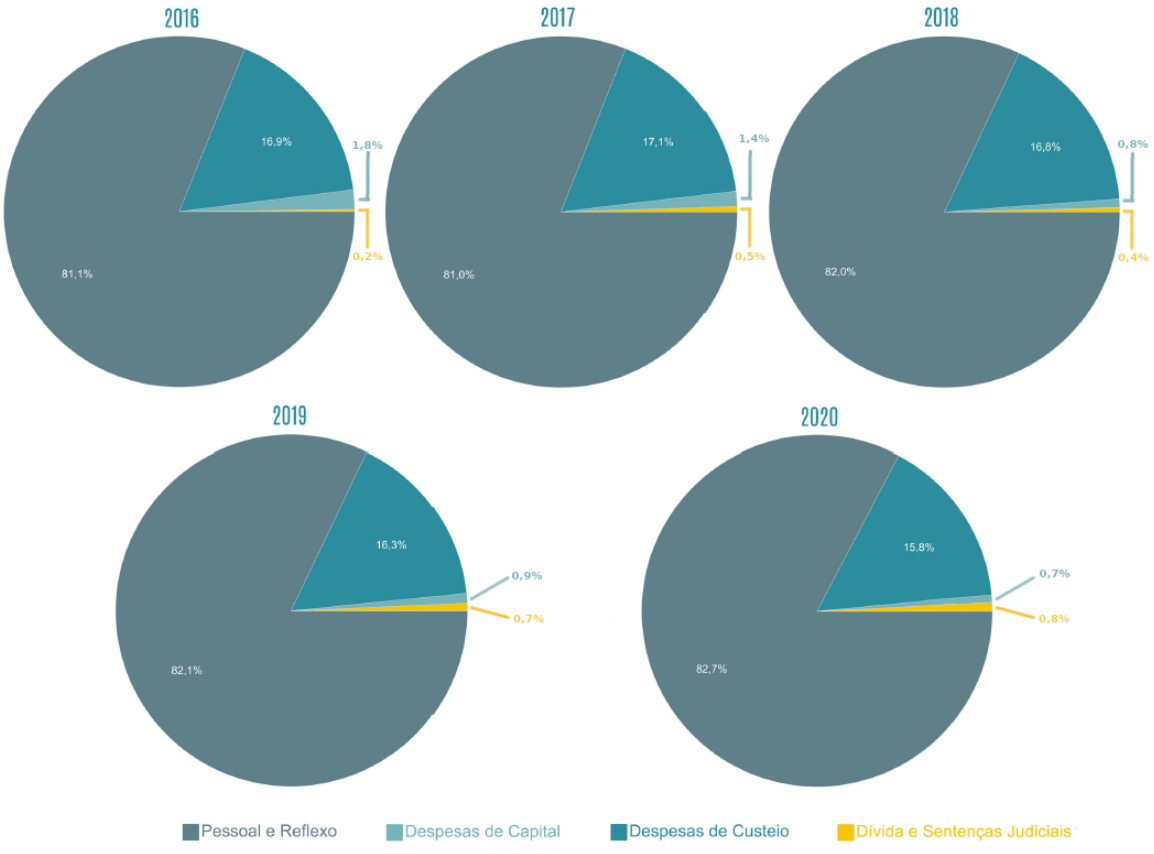
2.2 Contexto e Motivação

A forma de distribuição orçamentária na universidade, permite que a direção de cada unidade organize seus recursos de acordo com seu modelo de gestão. No modelo centralizado, a unidade administrativa é gerenciada como um único centro de custo da universidade. Ao passo que no modelo distribuído, a unidade pode ser subdividida em centros de custo hierarquicamente organizados. Contudo, independente do modelo de gestão, é necessário que o gestor compreenda

os recursos geridos por seu centro de custo.

A Figura 2 apresenta a distribuição dos gastos da Unesp nos últimos anos. Observa-se na figura que mais de 80% do orçamento da universidade foi empregado com despesas relacionadas à folha de pagamento, destacados nos gráficos como o item *"Pessoal e Reflexos"*. Nota-se também que uma pequena parte do orçamento é dedicada a compromissos judiciais, destacadas como *"Dívidas e Sentenças Judiciais"*. Sendo assim, pode-se constatar que a parte orçamentária dotada para gestão nas unidades, atualmente representa menos do que 19% do orçamento disponível, sendo esta parte composta pelos itens indicados nos gráficos como *"Despesas de Capital"* e *"Despesas de Custeio"*.

Figura 2 – Distribuição de gastos da universidade entre 2016 e 2020.



Fonte: Adaptado de APE (2021a).

Aprofundando a análise de despesas macro-orçamentárias da universidade, para o nível de unidade universitária, observa-se na Tabela 1 o detalhamento da análise financeira de uma das unidades da Unesp nos últimos anos. Os dados representados na tabela, foram obtidos do sistemas de gestão administrativa da Universidade, denominado SisADM, sistema de informação responsável por consolidar as transações financeiras e contábeis da Unesp. Este sistema está em funcionamento desde 2011, mas que teve sua utilização efetiva por todas as unidades universitárias apenas a partir do ano de 2018.

Tabela 1 – Análise financeira de registros da FEB no SisADM.

Ano	Descrição do Item Observado	Valor
2018	Valor total de despesas	R$ 30.927.726,34 (100%)
	Valor relacionado com folha de pagamento	R$ 27.059.758,76 (87,49%)
	Valor não relacionado com folha de pagamento	R$ 3.867.967,58 (12,51%)
2019	Valor total de despesas	R$ 36.518.504,84 (100%)
	Valor relacionado com folha de pagamento	R$ 29.781.743,54 (81,55%)
	Valor não relacionado com folha de pagamento	R$ 6.736.761,30 (18,45%)
2020	Valor total de despesas	R$ 34.269.754,57 (100%)
	Valor relacionado com folha de pagamento	R$ 29.373.828,46 (85,71%)
	Valor não relacionado com folha de pagamento	R$ 4.895.926,11 (14,29%)

Fonte: Produzida pelo autor.

Os dados da unidade universitária analisada nas Tabelas 1 e 2, são da Faculdade de Engenharia – Campus de Bauru (FEB). Unidade escolhida para detalhamento de informações e implantação experimental da solução desenvolvida neste trabalho, por ter como diretor da unidade durante o período de realização do estudo, o Prof. Dr. Luttgardes de Oliveira Neto, cliente demandante da solução e responsável por autorizar o desenvolvimento do estudo na faculdade. Também por ser a unidade de lotação do gerente de projetos responsável pelo sistema SisADM, o Sr. Rubens Memari Junior, que forneceu informações históricas relativas à implantação do sistema na Unesp e concedeu acesso aos dados necessários para o desenvolvimento deste trabalho.

Tabela 2 – Análise quantitativa de registros da FEB no SisADM.

Ano	Descrição do item observado	Valor
2018	Quantidade total de registros de despesas	2673 (100%)
	Quantidade de registros relacionados com folha de pagamento	307 (11,49%)
	Quantidade de demais registros, não relacionados com folha de pagamento	2366 (88,51%)
	Quantidade de registros com classificação contábil genérica (Outros...)	404 (17,08%)∗
2019	Quantidade total de registros de despesas	2521 (100%)
	Quantidade de registros relacionados com folha de pagamento	214 (8,49%)
	Quantidade de demais registros, não relacionados com folha de pagamento	2307 (91,51%)
	Quantidade de registros com classificação contábil genérica (Outros...)	406 (17,60%)∗
2020	Quantidade total de registros de despesas	1517 (100%)
	Quantidade de registros relacionados com folha de pagamento	169 (11,14%)
	Quantidade de demais registros, não relacionados com folha de pagamento	1348 (88,86%)
	Quantidade de registros com classificação contábil genérica (Outros...)	297 (22,03%)∗

∗ Valores percentuais relativos aos demais registros, não relacionados com folha de pagamento.

Fonte: Produzida pelo autor.

Considerando a grande quantidade de registros contábeis não relacionados com folha de pagamento apresentados na Tabela 2 (mais de 88%), considerando também que mais de 17% destes registros possuem classificação contábil genérica e pouco informativa como: *"Outros serviços..."*, *"Outros materiais..."*, *"Outras taxas..."*, etc., nota-se então a necessidade de discriminar este conjunto de informações genéricas, traduzindo-as em informações objetivas e de fácil entendimento para o gestor do centro de custo.

2.3 Problemas e Hipóteses

O principal problema abordado neste estudo é a elucidação de registros públicos contábeis. Tendo como foco mais específico, registros que de acordo com a legislação vigente são contabilmente categorizados de forma genérica. Ou seja, a elucidação de registros pouco informativos ao público de interesse. Um problema secundário abordado, é o estabelecimento de um processo para construção e integração de ferramentas de visualização em um sistema institucional em funcionamento.

A partir da questão inicial de como a Informática poderia contribuir para melhoria da compreensão humana sobre os dados financeiros e contábeis, foram estabelecidas quatro hipóteses iniciais que nortearam o desenvolvimento do trabalho:

- Hipótese 1 — Ferramentas de visualização da informação, criadas com informações padronizadas por ferramenta de classificação, são eficazes para transparência pública e eficientes para elucidação da execução orçamentária.
- Hipótese 2 — Técnicas e métodos da área de aprendizado de máquina, podem ser integradas ao processo de construção de ferramentas de visualização da informação.
- Hipótese 3 — Espaços de rótulos estruturados, podem ser utilizados como facilitadores na tarefa de classificação.
- Hipótese 4 — Códigos e descrições de elementos contábeis, podem ser utilizados como espaços de rótulos estruturados.

2.4 Fases do Estudo de Caso

Esta seção introduz as etapas do trabalho realizado, descrevendo principais atividades e resultados obtidos em cada fase do estudo.

2.4.1 Fase 1 – Diagnóstico

Fase inicial do estudo de caso, que contempla essencialmente a identificação do problema, formulação de hipóteses e o estabelecimento de objetivos para investigação.

A análise de registros financeiros e contábeis da faculdade estudada, demonstrou que estes apresentam informações diversificadas e de difícil entendimento para as chefias de departamentos. O contato com um modelo estático de visualização, criado manualmente a partir dos dados financeiros e contábeis, inspirou possíveis soluções para o problema identificado. Esta análise preliminar revelou a presença de problemas computacionais relacionados à Classificação de Texto, Processamento de Linguagem Natural e Visualização da Informação.

Desta forma, a Contabilidade e a Administração Pública foram identificadas como as principais áreas do conhecimento envolvidas com o problema exposto. Ao passo que o Aprendizado de Máquina e a Visualização da Informação, foram identificadas como as principais áreas do conhecimento envolvidas com as soluções idealizadas.

Como resultado da fase de diagnóstico elaborou-se um projeto de pesquisa (detalhado no Apêndice A), abordando o tema da criação e integração de ferramentas de visualização da informação aos sistemas institucionais legados. Nesta fase também foram identificados os *stakeholders* da faculdade estudada, ou seja, as pessoas e setores envolvidas e interessadas no contexto da pesquisa. Estabelecendo-se então um grupo de trabalho para atuar nas atividades deste estudo, grupo composto por quatro pessoas da área de gestão administrativa e três pessoas da área da tecnologia da informação. A Tabela 3 apresenta descrição do perfil funcional de cada membro e destaca algumas ações nas quais este se envolveu durante a realização do trabalho.

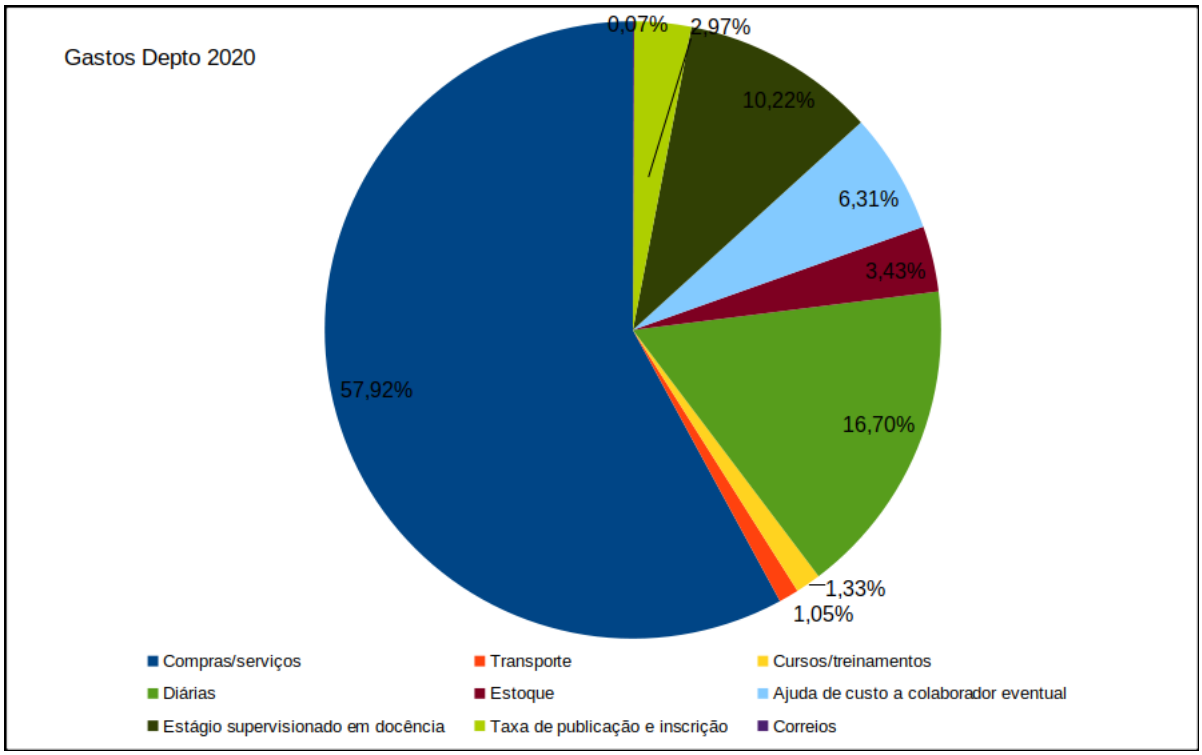
Tabela 3 – Descrição dos Membros do Grupo de Trabalho.

Cargo/Função	Descrição Profissional e Atuação no Estudo de Caso
Diretor de Unidade Universitária	Membro do corpo docente, responsável pela direção da faculdade no período de desenvolvimento do trabalho. Cliente demandante da solução e responsável pela autorização de execução do trabalho.
Diretor da Área de Gestão Administrativa	Membro do corpo técnico e administrativo, responsável pela divisão administrativa da faculdade. Cliente demandante da solução, atuando principalmente na validação de protótipos de visualização, validando dados de categorias e finalidades, também responsável pelo embasamento legal relacionado com assuntos financeiros e contábeis.
Assessor Administrativo	Membro do corpo técnico e administrativo, responsável por assessorar a diretoria administrativa. Pessoa responsável pela criação do modelo estático de visualização, participou da elaboração das classes de categorias e finalidades, realizou validação na classificação automática de dados.
Assistente Administrativo	Membro do corpo técnico e administrativo, responsável pelo setor de finanças da faculdade. Responsável pelo embasamento legal relacionado com assuntos financeiros e contábeis, atuou principalmente como validador de dados.
Diretor da Área de Tecnologia da Informação	Membro do corpo técnico e administrativo, responsável pela diretoria de informática da unidade. Cliente demandante da solução, colaboração na elaboração das classes de categorias e finalidades, atuou na classificação manual de dados e realizou validação na classificação automática de dados.
Analista de Informática	Membro do corpo técnico e administrativo, responsável pela gerência do projeto do sistema SisADM. Responsável por conceder acesso ao sistema institucional, também responsável pela validação de dados.
Assistente de Informática	Membro do corpo técnico e administrativo. Responsável pela análise e desenvolvimento da solução implementada, atuando também na classificação manual de dados e na elaboração das classes de categorias e finalidades.

Fonte: Produzida pelo autor.

A Figura 3 representa o modelo estático de visualização, criado manualmente utilizando-se software de planilha eletrônica. Para chegar neste resultado, um membro do grupo de trabalho categorizou os lançamentos contábeis registrados em planilha eletrônica e utilizou funcionalidade específica do software para geração de gráficos.

Figura 3 – Visualização criada manualmente utilizando software de planilha eletrônica.



Fonte: Produzida pelo autor.

2.4.2 Fase 2 – Fundamentação

A partir da identificação das disciplinas relacionadas tanto com o problema abordado, quanto com a solução idealizada, procedeu-se com o levantamento de referencial teórico sobre as áreas do conhecimento identificadas. Neste sentido, realizou-se uma Revisão Sistemática da Literatura (RSL) (detalhada no Apêndice B) sobre o tema Análise de Texto, com objetivo de identificar técnicas, métodos, aplicações, usos e contribuições do aprendizado de máquina para área de Visualização da Informação. Considerando especificamente qual é a contribuição ou correlação, de aprendizado de máquina com a visualização de informação textual, sob a perspectiva humana na relação de interação humano-computador.

A realização da RSL possibilitou a identificação de tópicos fundamentais da classificação de texto, do processamento de linguagem natural e da inteligência artificial (detalhados nos Apêndices D e E). Estabelecendo termos e conceitos relacionados aos temas introduzidos e contribuindo na identificação dos processos e procedimentos relacionados com a classificação de texto. Estabelecendo também, que a utilização de técnicas de aprendizado de máquina na análise de texto pode ser descrita pelas tarefas de mineração de texto, principalmente pela classificação automática de texto.

A RSL contribuiu para a descoberta de abordagens, métodos, técnicas e ferramentas aplicadas na tarefa de classificação automática de texto. Constatando-se que pela literatura não é possível determinar qual é a melhor técnica a ser utilizada para classificação, sendo necessário avaliar o problema e realizar experimentações para encontrar a abordagem mais adequada. Concluindo que estas abordagens podem ser utilizadas para estruturar documentos em corpus

para serem utilizados por técnicas de visualização da informação. Identificou-se também, que a aplicação de métodos de agrupamento ou classificação de documentos de texto, permitem a realização de análise visual do corpora.

Como resultado da revisão sistemática, estabeleceu-se que a contribuição do aprendizado de máquina para a visualização da informação, está relacionada com o pré-processamento dos dados a serem utilizados pelas ferramentas de visualização. Concluindo-se que é viável a aplicação conjunta de métodos e técnicas das áreas aprendizado de máquina e visualização da informação, para construção de sistemas de inteligência comercial (BI, do inglês *Business Intelligence*) e sistemas de análises visuais (em inglês, *Visual Analytics*). Sistemas estes, que possibilitam trabalhar os dados dos registros contábeis, de forma mais amigável e menos burocráticas para o usuário humano, permitindo a exploração das informações persistidas nos registros financeiros e contábeis.

2.4.3 Fase 3 – Execução

O conhecimento consolidado na fase de fundamentação, sugere que o problema motivador deste trabalho poderia ser abordado utilizando-se de forma conjunta a classificação automática de texto e o processamento de linguagem natural, para padronização de informações dos registros financeiros e contábeis. Sendo a classificação de categorias um problema de classificação hierárquica de texto, ao passo que na identificação de finalidade tem-se um problema de classificação multiclasse plano, ambos podendo ser abordados com adaptações de técnicas semânticas, como o uso de dicionários léxicos.

◆ Decisões de Projeto

A fase de execução do estudo de caso, iniciou-se com a tomada de decisões pelo grupo de trabalho estabelecido para o estudo. Destacam-se duas decisões de projeto que impactaram diretamente no desenvolvimento do trabalho:

- Aquisição ou Implementação?

  Optou-se pela implementação total da solução, para consolidação dos conceitos estabelecidos e perpetuação deste conhecimento na equipe de tecnologia da informação.

- Solução Integrada ou Ferramenta Externa?

  Optou-se por realizar a implementação da solução de forma integrada ao SisADM, aproveitando-se de fontes de informações fornecidas por serviços deste sistema. Contudo, a solução implementada foi desenvolvida de forma modularizada e não invasiva, podendo facilmente ser desacoplada do SisADM sem prejuízo de suas funcionalidades.

◆ Construção da Solução

Estabelecidas as diretrizes de trabalho, foram desenvolvidas as atividades para construção da solução:

- Aquisição e Armazenamento de Dados — Primeiro, foram estabelecidas as fontes de dados e a forma de armazenamento de informações para serem utilizadas nas ferramentas de visualização.

  Os dados dos registros financeiros e contábeis foram obtidos de duas entidades do sistema SisADM. Entidade *"Receita"*, responsável por armazenar informações sobre lançamentos de crédito. Entidade *"Despesa"*, responsável pelos lançamentos do tipo débito. Todos os dados foram armazenados em um *Data Mart*[1] criado especificamente para a solução de prestação de contas e exploração da execução orçamentária.

- Padronização de Dados — O grupo de trabalho estabeleceu cinco finalidades para serem atribuídas aos registros financeiros e contábeis, refletindo as atividades-fim e atividades meio da instituição. Também foram estabelecidas 134 categorias para tipificação discriminativa dos registros, sendo 12 categorias do tipo crédito e 122 do tipo débito. Os nomes de categorias e finalidades foram padronizados na forma curta, sendo representados por uma única palavra.

- Categorização de Registros — Um conjunto de dados experimental foi manualmente classificado por membros do grupo de trabalho. A partir desta classificação, foram estabelecidas as palavras-chave relacionadas com cada classe de Categoria e Finalidade estabelecidas. Procedeu-se então com o desenvolvimento de um algoritmo de classificação de texto, um método de classificação autoral que utiliza a abordagem lexical e é baseado em palavras-chave, método utilizado para realizar a classificação automática dos registros do *Data Mart*, atribuindo-lhes categorias discriminativas e finalidades qualitativas.

- Protótipos da Ferramentas de Visualização — Foram elaborados protótipos de ferramentas estáticas e interativas para a visualização de informações de interesse. Inicialmente um gráfico interativo para navegação em profundidade, para exploração de registros por Categoria. Além de gráficos estáticos para apresentação da distribuição de recursos entre Finalidades e centros de custo.

- Protótipo da Solução — A partir dos protótipos de ferramentas visuais validados, foram desenvolvidos painéis de visualização compostos por ferramentas distintas, para apresentação de informações contextualmente relacionadas.

- Integração da Solução — A solução desenvolvida foi integrada ao sistema SisADM para obtenção e classificação automática dos registros financeiros e contábeis. Os registros classificados automaticamente foram disponibilizados em uma ferramenta específica de prestação de contas, onde os gestores dos centros de custo poderiam confirmar ou ajustar a classificação dos registros do *Data Mart*.

- Divulgação e Avaliação da Solução — Por fim, foram realizadas sessões de apresentação da solução e sessões de treinamento dos gestores dos centros de custo. Nas sessões de treinamento, a ferramenta de prestação de contas e as ferramentas para exploração da

[1] Um *Data Mart* é um tipo simples de Data Warehouse, focado em um único assunto ou linha de negócios. (ORACLE, 2023, tradução livre)

execução orçamentária, foram apresentadas com riqueza de detalhes e propondo-se a realização de atividades práticas. Estas sessões também foram utilizadas para se obter a avaliação qualitativa da solução desenvolvida, avaliação realizada pelo público-alvo relacionado.

2.5 Considerações Finais sobre o Estudo de Caso

Este capítulo descreveu brevemente o estudo de caso realizado neste trabalho. Nos capítulos seguintes serão apresentados o referencial teórico, obtido na fase de fundamentação, o algoritmo de classificação e as ferramentas de visualização da informação, desenvolvidos na fase de execução do estudo.

3 Referencial Teórico

Este capítulo apresenta conceitos fundamentais sobre temas das áreas do conhecimento relacionadas com este trabalho.

3.1 *Visual Analytics*

Tufte (2006) estabelece os princípios fundamentais de um projeto analítico, que deve integrar evidências com palavras, números e imagens, possibilitando uma análise multivariada de dados, permitindo comparações, mostrando contrastes e diferenças. O autor destaca que o conteúdo é mais importante que tudo.

A análise visual de dados é realizada para encontrar-se padrões previamente desconhecidos ou que se destacam da norma (*outliers*). O cérebro visual humano é um poderoso mecanismo de busca de padrões, sendo esta a razão fundamental pela qual técnicas de visualização estão se tornando importantes. Métodos de visualização são a melhor maneira de apresentar informações, para que seja possível a descoberta de estruturas, grupos e tendências registradas em centenas de valores de dados (MUNZNER, 2014).

Zayas et al. (2017) conceitualizam *analytics* como um processo de análise sistemática de dados, que utiliza várias técnicas para obter *insights* de um conjunto de dados. Os autores comentam que as técnicas de *analytics* são baseadas na combinação de regras de negócios, análises estatísticas, algoritmos, análise de texto, visualização da informação e outros. Comentam ainda que o desenvolvimento de uma plataforma de análise visual é custoso em relação a recursos humanos, financeiros e de tempo, mas que as informações disponibilizadas através deste tipo de plataforma vão além do conceito de inteligência comercial, culminando com a transformação da empresa e seus processos.

3.2 Visão Geral sobre Visualização da Informação

A Visualização da Informação é a área do conhecimento dedicada ao estudo de sistemas de visualização baseados em computador. Sistemas estes, que promovem representações visuais de conjuntos de dados, com objetivo de auxiliar pessoas a realizarem tarefas de forma mais eficientes (MUNZNER, 2014). Segundo o autor, é enorme o escopo de estudo da área de visualização da informação, contemplando principalmente considerações de como criar e como interagir com representações visuais. Contemplando também o estudo do design, do processo, da construção e da avaliação de ferramentas de visualização. Nota-se na definição apresentada a presença de pessoas e computadores, sendo o elemento humano o principal ator em um sistema de visualização de informações.

3.2.1 Interação Humano-Computador na Visualização de Informações

Soluções de visualização usufruem da significativa capacidade de processamento de informação visual que o cérebro humano possui. O sistema visual humano possui uma ligação com cérebro através de um canal com altíssima largura de banda, provendo a transmissão de significativa quantidade de informações visuais, cujo processamento ocorre de forma paralela e em nível pré-consciente (MUNZNER, 2014).

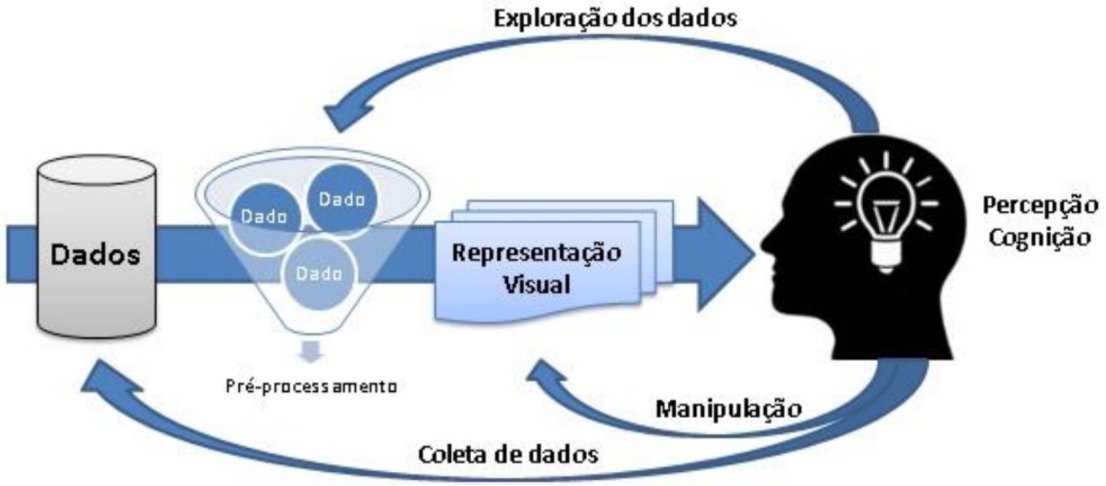
Ware (2012) descreve o processo de visualização da informação em quatro etapas, que contemplam desde a origem dos dados até o processamento da informação pelo cérebro humano:

- Coleta e Armazenamento de Dados — Etapa de obtenção e registro dos dados de interesse.
- Pré-processamento — Responsável pela sanitização dos dados, com objetivo de torná-los mais amigáveis para manipulação e compreensão.
- Representação Visual — Projeção dos dados em ferramentas de visualização da informação suportadas por computador.
- Percepção do Significado da Representação — O processo cognitivo humano.

A Figura 4 representa o processo descrito por Ware (2012). O autor descreve que é comum que sejam realizados na etapa de pré-processamento, procedimentos como: i) a resolução de ambiguidades textuais; ii) a redução da quantidade de dados; e, iii) a transformação e padronização de dados tais como nomenclaturas ou unidades métricas quantitativas.

Com relação à representação visual, Munzner (2014) destaca que ferramentas de visualização são representações externas que possibilitam estender a memória interna e melhorar a capacidade de cognição humana. O autor indica que uma das vantagens da utilização de gráficos como memória externa, é a possibilidade de organizar informações por localização espacial, o que acelera tanto a busca quanto o reconhecimento de padrões.

Figura 4 – Etapas do Processo de Visualização da Informação.



Fonte: Extraído de (DA SILVA RODRIGUES; BREGA, 2017).

Munzner (2014) também ressalta a importância de se considerar as limitações envolvidas nesta relação de interação humano-computador na visualização de informações. O autor elenca três principais tipos de limitações que afetam de diferentes formas os componentes da relação:

- Capacidade Perceptual e Cognitiva Humana — A atenção do ser humano pode ser facilmente prejudicada. E a capacidade da memória humana para o armazenamento de longo prazo de informações não visuais é limitada. Por fim, surpreendentemente, a memória de trabalho visual (memória de curto prazo) armazena poucas informações pode nos deixar vulneráveis à cegueira da mudança.
- Capacidade Computacional — Recursos computacionais como a memória e a capacidade de processamento, são recursos limitados e finitos. O tamanho do conjunto de dados pode ultrapassar a capacidade de memória do computador. Da mesma forma, é uma grande preocupação a complexidade computacional dos algoritmos para pré-processamento de dados, transformação, layout e renderização.
- Capacidade de Exibição. — Por vezes, o tamanho e a resolução da tela disponível, não é suficiente para mostrar simultaneamente todas as informações desejadas. Deste modo, o projetista de ferramenta de visualização deve ponderar entre os benefícios de apresentar tudo que for possível de uma única vez, o que minimiza a necessidade de navegação e exploração, contrapondo com os malefícios de apresentar muitas informações, o que pode prejudicar a experiência do usuário pela desordem visual.

3.2.2 Motivações para Construção de Ferramentas de Visualização

Ferramentas de visualização podem ser projetadas para diversas finalidades. Estas ferramentas ajudam pessoas analisarem a estrutura do conjunto de dados, seja de forma exploratória para se encontrar padrões, confirmando os esperados ou descobrindo os inesperados, ou seja como ferramenta de apoio na avaliação e análise modelos estatísticos, julgando-se a adequação do modelo aos dados. Ferramentas de visualização permitem que pessoas analisem dados para encontrar respostas a perguntas que não sabiam que deveriam ser feitas (MUNZNER, 2014).

O autor destaca que uma solução de visualização é adequada quando existe a necessidade de aumentar as capacidades do agente humano envolvido no processo de tomada de decisão, ao invés de simplesmente substituir as pessoas por métodos computacionais de tomada de decisão. Destacando ainda, o fato da necessidade de julgamento humano sobre as informações relacionadas, ser fator determinante para criação (ou não) de uma solução de visualização de informações. Pois se existe solução aceitável baseada em computador que consiga tomar decisões de forma completamente automatizada, então não existe necessidade de se projetar uma ferramenta de visualização.

3.2.3 Motivações para Utilização de Ferramentas de Visualização

A utilização de ferramentas de visualização na análise de dados, pode ser feita por pessoas que desejam apenas consumir informações existentes, ou pode ser realizada por usuários

interessados em produzir ativamente novas informações. Sendo mais comum o caso dos usuários que desejam apenas consumir informações existentes (MUNZNER, 2014). Neste caso, o autor informa que a utilização de ferramentas de visualização são motivadas por três principais objetivos:

- Descobrir — A meta de descoberta consiste na utilização da visualização para encontrar novos conhecimentos. Esta utilização pode ser com objetivo de encontrar coisas completamente novas ou de descobrir se uma conjectura é verdadeira ou falsa.
- Apresentar — Utilização da visualização para comunicação sucinta de informações com *storytelling*, que consiste na contação de história com apresentação de dados para orientar um público através de uma série de operações cognitivas. A principal característica deste objetivo é a utilização da visualização para comunicar algo específico e já compreendido para o público, podendo ocorrer em contextos institucionais como a tomada de decisão, planejamento, previsão e processos de instrução.
- Apreciar — A apreciação refere-se a encontros casuais com visualizações, por exemplo, ao se observar um infográfico que acompanha um texto publicado. Neste sentido, a utilização de ferramentas de visualização não é motivada por uma necessidade, mas pela curiosidade. Curiosidade esta que pode ser satisfeita ou estimulada pela ferramenta de visualização.

3.2.4 Diretrizes para Projeto de Construção de Ferramentas de Visualização

Segundo Munzner (2014), o projeto de uma ferramenta de visualização deve ser balizado pela tarefa objetivo do usuário, ou seja, é necessário pensar no que o usuário pretende fazer para que a ferramenta seja adequada à tarefa. Pois uma mesma ferramenta de visualização pode servir bem para uma tarefa e ao mesmo tempo ser inadequada para outra.

A criação e manipulação de representações visuais segue o idioma de visualização estabelecido para tal, sendo possível criar de muitas maneiras uma representação visual de dados como uma única imagem. Representações estáticas simples podem ser concebidas com vários tipos de gráficos (e.g., gráfico de barras, gráfico de linhas, gráfico de pizza, gráfico de rosca). E as possibilidades de design expandem-se ao considerar a manipulação de uma ou mais dessas imagens de forma interativa, de modo que a união de múltiplos diagramas simples por meio da interação apresenta um idioma de visualização mais complexo.

A interatividade é crucial em ferramentas de visualização que lidam com complexidades, pois permitem a investigação em vários níveis de detalhe, desde uma macrovisão resumida até uma microvisão detalhada, permitindo também representar relações, conexões e dependências de informações (e.g., hierarquias, especializações, generalizações).

Tufte (2001) discorre sobre a representação visual da informação quantitativa, mostrando como utilizar pontos, linhas, sistema de coordenadas, números, símbolos, palavras, sombreamento e cores, para apresentar visualmente quantidades medidas. O autor estabelece que bons gráficos servem a um propósito claro, devendo mostrar os dados. A ferramenta de visualização deve induzir o usuário a pensar sobre o conteúdo e não sobre a forma. Bons gráficos são capazes de representar grandes conjuntos de informação de forma coerente e evitando distorções. Estes

gráficos devem revelar dados em vários níveis de detalhamento e deve ser possível a comparação entre trechos de dados.

Kirk (2012) fornece orientações para que projetistas de visualizações tenham sucesso na construção de ferramentas de visualização de informações. O autor discorre, por exemplo, sobre objetivos típicos de visualizações e realizada a indicação de ferramentas adequadas para se atingir estes objetivos. A Tabela 4 apresenta exemplos da relação estabelecida pelo autor entre os objetivos típicos e as ferramentas indicadas.

Tabela 4 – Relação de Objetivos com Métodos de Visualização de Dados.

Finalidade da Visualização	Método de Visualização Indicado
Representação de relações hierárquicas Representação de relação parte-de-todo	Gráfico de Pizza/Rosca
Comparação monovalorada de categorias	Gráfico de Barras
Comparação multivalorada de categorias	Diagrama de Sankey
Representação de mudanças ao longo do tempo	Gráfico de Linha

Fonte: Produzida pelo autor.

3.2.5 Ferramentas de Visualização

Esta seção dedica-se à descrição das ferramentas de visualização implementadas neste trabalho.

◆ Gráfico de Pizza e Gráfico de Rosca

Os gráficos de pizza e rosca são representados por figura circular, sendo o gráfico de pizza um círculo com a totalidade de sua área preenchida pelo conteúdo, enquanto o gráfico de rosca possui a região central sem conteúdo, permitindo a inclusão de rótulos ou ícones nesta região.

Segundo Kirk (2012), os gráficos circulares são provavelmente o tipo de gráfico mais controverso, pois pode atrair um sentimento negativo devido a dificuldade humana em interpretar com precisão ângulos e avaliar (ou comparar) áreas de segmentos circulares. O autor também destaca que esta negatividade é na realidade um reflexo do implacável mau uso desta ferramenta. Desta forma, não é indicada a inclusão de muitas categorias e cores, também sendo fortemente contraindicada a utilização de decorações 3D. Recomenda-se sempre iniciar o primeiro corte a partir da posição vertical, para melhor disposição do arranjo visual da ferramenta.

◆ Gráfico de Barras

A ferramenta gráfico de barras, que também pode ser chamada de gráfico de colunas, transmite informações com a utilização da altura ou da largura de uma barra. A disposição das barras lado a lado possibilita ao usuário realizar comparações precisas entre valores relativos ou absolutos das categorias apresentadas.

É importante mostrar toda a extensão da propriedade quando se utiliza o comprimento como a variável visual para representar um valor quantitativo, por isso, sempre deve-se iniciar

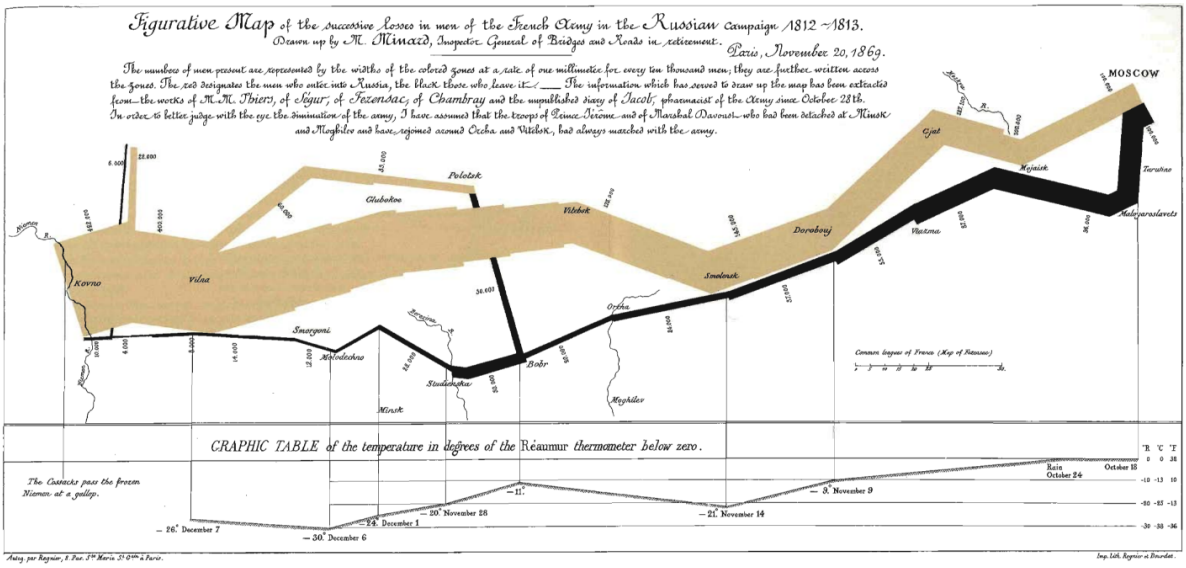
a barra do ponto zero do eixo. E o uso da coloração pode ajudar no destaque de categorias específicas e na construção da narrativa relacionada com a informação (KIRK, 2012).

◆ Diagrama de Sankey

O Diagrama de Sankey é uma ferramenta de visualização que permite representar o fluxo de um conjunto de valores. Os elementos terminais do fluxo são chamados de nós (*node*, em inglês), sendo estes conectados pelos elementos chamados de *links*. Este diagrama é indicado para demonstrar mapeamentos de muitos para muitos entre dois domínios, sendo também indicado para representar vários caminhos em um evento separado por etapas (KIRK, 2012; GOOGLE, 2022).

Um famoso exemplo na área de visualização de informação, para a utilização deste tipo de diagrama representando fases de um evento, é o trabalho de Charles Joseph Minard representado na Figura 5, na qual estão demonstradas as baixas do exército de Napoleão, durante a campanha contra a Rússia nos anos de 1812 e 1813.

Figura 5 – Representação de Charles Joseph Minard para as baixas no exército de Napoleão.



Fonte: Tufte (2006, pág. 123-124).

No Diagrama de Sankey, nós e *links* são representados de acordo com a variação de valores atribuídos. Ou seja, quanto maior o valor, maior será o tamanho do elemento relacionado. Segundo o Google (2022), na documentação de sua plataforma *Google Charts*[1], este tipo de diagrama recebeu o nome do capitão Sankey, que criou um diagrama de eficiência do motor a vapor, utilizando setas com larguras proporcionais à perda de calor.

◆ Gráfico de Linha

As ferramentas do tipo gráfico de linhas são provavelmente algo que a maioria das pessoas estejam familiarizadas. Este tipo de gráfico é utilizado para analisar uma variável quantitativa

[1] https://developers.google.com/chart/interactive/docs

contínua no eixo X, registrando-se os valores periodicamente medidos no eixo Y. Os pontos de medição são unidos usando-se linhas que demonstram a trajetória de tendência (descendente, estável, ascendente) observada na medição temporal. Destaca-se que ao contrário dos gráficos de barras, o eixo Y não precisa começar no valor zero, pois a ferramenta apresenta um padrão relativo do percurso de dados (KIRK, 2012).

Segundo Castillo (2014), gráficos de linha são ideais para representação de séries temporais, pois permitem a fácil identificação de pontos sazonais discrepantes. Este conceito frequentemente é expresso em inglês como *outliers*. O autor destaca ainda que caso seja necessário precisão de valores, a representação em tabela é a melhor escolha, mas se a exatidão não for necessária, o gráfico de linhas é uma excelente ferramenta para representação de séries temporais.

Para Munzner (2014), séries temporais são a forma de representação mais utilizada de design gráfico. Séries temporais são utilizadas para apresentar valores medidos em intervalos regulares de tempo. Os registros de preços de uma ação ou os registros de batidas do coração em um eletrocardiograma, são bons exemplos de séries temporais.

◆ *Dashboards*

A palavra *dashboard* poderia ser traduzida para Língua Portuguesa como "painel", no sentido de ser um "painel visual". Few (2006, tradução livre) define *dashboard* como *"uma exibição visual das informações mais importantes necessárias para atingir um ou mais objetivos; consolidados e organizados em uma única tela para que as informações possam ser monitoradas rapidamente"*.

Castillo (2014) descreve que *dashboards* são exibições densas de gráficos para ajudar a entender da maneira mais rápida e eficaz possível, as principais métricas de um problema observado. Sendo esta ferramenta frequentemente utilizada em soluções de BI, para apresentação de painéis de inteligência de negócios. O autor destaca ainda que a principal característica de um *dasboard* é que toda informação possível deve estar visível e disponível para o consumo imediato do usuário, de modo que toda a informação seja entregue rapidamente.

3.2.6 Avaliação de Ferramentas de Visualização

Segundo Kirk (2012), projetistas de soluções de visualização podem conhecer grandes conceitos e ter impressionantemente ideias criativas, que podem ser concretizadas utilizando-se os recursos tecnológicos adequados e disponíveis. Após esta concretização e lançamento da ferramenta, o projetista deve avaliar a eficácia e o impacto das visualizações, buscando identificar o quão bem seu projeto serviu ao propósito de criação determinado no início do processo.

O autor elenca que os principais *feedbacks* de interesse são:

- Houve uma reação positiva à peça criada?
- A solução ofereceu o tom apropriado?
- O trabalho atingiu o tipo e volume de público de interesse?

- Os usuários foram capazes de consumir ou descobrir ideias de forma eficaz?
- Quais os problemas que as pessoas experimentam, se houveram?

Para se obter estes *feedbacks* de interesse, o autor sugere alguns possíveis métodos de avaliação, por exemplo:

- Métricas e Indicadores de Referência — Quantificação de informações estatísticas, utilizada frequentemente em soluções *Web* para indicar o alcance e popularidade da página.
- *Feedback* de Cliente — Retorno daqueles que demandaram a visualização, que obviamente possuem a opinião mais relevante sobre a solução apresentada.
- Revisão por Pares — Avaliações importantes e construtivas podem ser realizadas por profissionais especialistas, estudiosos ou líderes de pensamento.
- *Feedback* Não Estruturado — Valor obtido por comentários espontâneos, frequentemente realizados em soluções online que permitem esta manifestação do usuário.
- Solicitação de Avaliação — Convite realizado de forma pró-ativa para que os usuários avaliem a solução proposta.
- Estudo de Caso Formal — Nível avançado de avaliação, com utilização de método científico.

3.3 Visão Geral sobre Classificação de Texto

Numa definição ampla, a classificação de texto é o processo de atribuir categorias à amostras de texto. Este conceito de também pode ser encontrada na literatura como categorização de texto (KUMAR; RAVI, 2016; RAZA et al., 2019). A tarefa de classificação de documentos, é uma especificação da classificação de texto, também consistindo na atribuição de uma ou mais classes a um documento de texto (KOTU; DESHPANDE, 2019). Um tradicional exemplo de aplicação desta tarefa é a classificação de emails, realizada para identificação de mensagens de *spam* e *phishing*, ou realizada para categorização de mensagens por tópicos de interesse.

Maron (1961) e Borko e Bernick (1963) demonstram que a pesquisa em classificação automática de texto não é recente, remontando à década de 60. Mas este é um tema de pesquisa que ainda demonstra-se atual, tendo como principal foco a melhoria nos métodos ou abordagens de classificação. Analisando-se o tópico Classificação de Documentos da ScienceDirect (2023), percebe-se que este é um tema de pesquisa multidisciplinar. Tema relacionado por exemplo, mas não exclusivamente, com a Ciência da Informação (MAALEJ et al., 2016), com a Ciência da Computação (WALLACE; PAVLENKO, 2011) e com o Processamento de Linguagem Natural (YOUNG et al., 2018).

Guo, Shi e Tu (2016) descrevem que é possível distinguir abordagens de análise de texto em três categorias: Lexical, Semântica e *Machine Learning*. A abordagem lexical se refere a técnicas relacionadas com medidas de legibilidade de texto como *Fog Index* ou, refere-se a técnicas baseados em dicionários léxicos como *Harvard General Inquirer (Harvard GI)* ou *Loughran and*

Mcdonald dictionary (LM dictionary). Os autores destacam o trabalho de Ingram e Frazier (1980) como o primeiro trabalho a introduzir a análise de conteúdo textual, utilizando-se da contagem de frequência de palavras-chave para realizar esta abordagem lexical. A abordagem semântica dedica-se à extração de conteúdo conceitual em documentos de texto, buscando também a identificação de relações entre documentos. E a abordagem *Machine Learning* é empregada em diversas tarefas de mineração de texto.

Tendências recentes apresentam utilização de técnicas cada vez mais complexas para realização de tarefas de classificação de texto, com pesquisadores utilizando abordagens de *Machine Learning* (ML), *Deep Learning* (DL) e Inteligência Artificial (AI, do inglês *Artificial Intelligence*), para obterem melhores resultados de classificação em problemas específicos (HINGMIRE et al., 2013; YOUNG et al., 2018; MIROŃCZUK; PROTASIEWICZ, 2018; PIKIES; ALI, 2019; IBRAHIM et al., 2021; MA et al., 2022). Apesar de toda evolução nos métodos de classificação de texto, nem todo problema de classificação necessita da utilização de algoritmos complexos e de difícil implementação (*Hard Computing solution*). A teoria *Soft Computing* denota que problemas suaves podem ser enfrentados por técnicas tradicionais (ZADEH, 1994; IBRAHIM, 2016).

3.4 Tipos de Problemas de Classificação de Texto

Embora a conceitualização ampla da classificação de texto ser bastante simples, esta definição varia de acordo com o modo operante e com o problema de classificação abordado. No quesito modo operante, a classificação pode ser automática ou manual. Sendo a classificação automática de texto a mais proeminente dentre todas as tarefas da mineração de texto, tanto que 67% dos trabalhos analisados no estudo de Amani e Fadlalla (2017) estavam relacionados com a tarefa de classificação de texto. Quanto aos problemas de classificação, eles podem ser divididos em: i) Planos ou Hierárquicos; ii) Binários ou Multiclasse; e, iii) Monorrótulo ou Multirrótulo.

A nomenclatura dos tipos de problemas é definida por características específicas de cada problema, podendo um mesmo problema se enquadrar em mais de um tipo. Os tópicos a seguir descrevem os seis principais tipos de problemas de classificação.

3.4.1 Classificação Binária

Modelo de problema onde o conjunto de soluções é de ordem binária. O conjunto de classes possíveis de serem atribuídas às amostras consiste em apenas duas opções.

Exemplos de classes para problemas binários: {0,1}, {é, não é}, {verdadeiro, falso} e {crédito, débito}.

3.4.2 Classificação Multiclasse

Neste modelo, o conjunto de soluções tem um tamanho finito variando de *três* até n.

Exemplos de classes para problema multiclasse: {auxílios, compras, diárias, transportes}.

3.4.3 Classificação Monorrótulo

Abordagem relacionada aos problemas multiclasse, normalmente com categorias mutuamente excludentes. É um problema onde o domínio da aplicação aceita a atribuição de uma e somente uma categoria para cada amostra observada.

Exemplos de categorias para problema monorrótulo: {pessoa, veículo, planta}.

3.4.4 Classificação Multirrótulo

Abordagem também relacionada aos problemas multiclasse. Mas neste caso aceita-se a atribuição de várias categorias para cada amostra observada.

Exemplo: Utilizando as categorias {ensino, pesquisa, extensão universitária}, um artigo pode ser categorizado como atividade relacionada ao ensino e a pesquisa simultaneamente.

3.4.5 Classificação Plana

Problema de classificação onde o conjunto de classes possíveis de serem atribuídas às amostras, não apresenta relação de generalização ou especialização. Ou seja, o conjunto de classes é formado por categorias independentes umas das outras.

Exemplos de categorias para classificação plana: {ensino, pesquisa, extensão universitária}.

3.4.6 Classificação Hierárquica

Modelo que apresenta relações de especialização ou generalização entre as classes do conjunto de categorias possíveis.

Exemplos de classes hierárquicas: {despesas, despesas.auxilio, despesas.auxilio.pesquisador, despesas.auxilio.estudante, despesas.auxilio.colaboradoreventual}. Nota-se nestes exemplos a utilização do caractere ponto para realizar a concatenação de palavras e representar a profundidade hierárquica.

3.5 Abordagens para Problemas de Classificação de Texto

Problemas de classificação com diferentes características, apresentam diferentes abordagens de enfrentamento. Sendo estas abordagens relacionadas diretamente com a estrutura do problema ou com a quantidade de rótulos que poderá ser atribuído às amostras.

3.5.1 Abordagens para Problemas de Classificação Monorrótulo

Quando o conjunto de classes possíveis de serem atribuídas é formado por classes mutualmente excludentes, basicamente duas abordagens são adotadas para classificação (TSOUMAKAS; KATAKIS, 2007; READ et al., 2011; SCHRÖDER, 2018):

- *One-vs-All* — Um-contra-Todos é um método que cria K classificadores, um para cada classe do conjunto, então as amostras são submetidas a todos estes classificadores. Aquele que obtiver a menor diferença (maior semelhança) é a classe que atribuirá rótulo à amostra.

- *One-vs-One* — Na abordagem Um-contra-Um o problema multiclasse é transformado em vários problemas de classificações binárias. Considerando um conjunto de classes K, serão gerados $K(K-1)/2$ classificadores binários para realização de comparações de duas em duas classes. O classificador que ganhar mais embates 1×1 será a classe atribuída à amostra.

3.5.2 Abordagens para Problemas de Classificação Multirrótulo

Quando as classes do conjunto de categorias não são mutualmente excludentes, ou seja, a amostra pode ser rotulada com várias categorias, Metz (2011) e Schröder (2018) apresentam três abordagens possíveis:

- *Binary Relevance (BR)* — Na Relevância Binária, constrói-se um classificador discriminante (*é* ou *não é*) para cada rótulo possível. A amostra será analisada por todos estes classificadores e serão atribuídos os rótulos considerados relevantes (*é*).
- *Label Powerset (LP)* — Esta abordagem consiste na transformação do problema multiclasse-multirrótulo em problema multiclasse-monorrótulo. Criando-se novos rótulos que são compostos pelas combinações dos rótulos possíveis no conjunto de treinamento.
- *Stacking* — A ideia do Empilhamento é a utilização de metaclassificadores, onde os de mais alto nível utilizam como entrada a saída gerada pelos metaclassificadores base.

3.5.3 Abordagens para Problemas de Classificação Hierárquica

Segundo Freitas e Carvalho (2007) e Metz (2011), o problema da classificação hierárquica de texto, pode ser desenvolvido utilizando-se duas abordagens de classificação:

- *Mandatory Leaf Node Prediction (MLN)* — A Predição Obrigatória em Nós-Folha, é a abordagem que considera somente classes alocadas em nós terminais, como elegíveis para atribuição às amostras.
- *Non-mandatory Leaf Node Prediction (NMLN)* – A Predição Opcional em Nós-Folha, permite que qualquer classe da estrutura hierárquica seja atribuída às amostras observadas.

3.6 Visão Geral sobre o Processo de Classificação de Texto

Esta seção apresenta o processo de classificação de texto pela perspectiva da área de mineração de texto, considerando a classificação automática de texto apoiada por métodos tradicionais de aprendizado de máquina. Apesar desta não ser a abordagem adotada neste trabalho, alguns conceitos descritos aqui são fundamentais para o melhor entendimento deste.

Considerando tanto a visão mais simplificada do processo de classificação exposto em Kumar e Ravi (2016), quanto a visão mais sofisticada e detalhada do processo proposta em Mirończuk e Protasiewicz (2018), é possível agrupar as etapas do processo de classificação de texto em três grandes fases: pré-processamento; treinamento; e, validação. Embora vários

procedimentos serem elencados nas etapas do processo de classificação de texto, não existe a obrigatoriedade de execução de todos os procedimentos.

3.6.1 Pré-Processamento

Na fase de pré-processamento são realizadas análises morfológicas, semânticas e sintáticas das palavras que compõe o texto. A realização deste conjunto de análises culmina em duas grandes ações que são realizadas nesta fase: i) modelagem de documento; e, ii) seleção de características e tratamento de dimensionalidade. Considerando-se diferentes perspectivas sobre o processo de classificação, é consenso que a fase de pré-processamento é fundamental para qualquer solução de mineração de texto, pois ações realizadas neste ponto do processo influenciam diretamente nos resultados finais (KUMAR; RAVI, 2016; MEDEIROS, 2018; MIROŃCZUK; PROTASIEWICZ, 2018; RAZA et al., 2019; SANTOS; MERSCHMANN, 2020).

- *Modelagem de Documento* — A modelagem de documento tem como objetivo a transformação do documento de texto original, em um modelo que seja adequado para ser processado por algoritmos computacionais.

  Procedimentos executados:

  - *Case Folding* – Padronização de texto em caixa baixa ou alta (frequentemente em baixa).
  - *Cleaning* – Remoção de caracteres especiais, dígitos, sinais de pontuação e acentuação.
  - *Stopwords Removal* – Remoção de palavras irrelevantes como artigos e preposições.
  - *Length Filtering* – Remoção de palavras pequenas (normalmente menor que três caracteres).
  - *Tokenization* – Transformação de palavras em tokens *n-grams*, onde n indica a quantidade máxima de palavras que formam um token (*1-gram* é o mais comum, mas encontra-se usos de *2-grams* e *3-grams*).
  - *Stemming* – Redução de palavras ao seu radical.
  - *Lemmatisation* – Alteração de formas flexionadas de palavras, para uma forma padrão.
  - *Document Representation* – Construção da matriz de termos do documento e definição do modelo de representação (vetorial, grafos, *part-of-speech*, etc.).

- Extração de Características e Tratamento de Dimensionalidade — Parte mais importante do pré-processamento, consequentemente também é considera parte mais importante da mineração de texto. Na literatura existem muitos estudos que concentram-se apenas neste tópico.

  Procedimentos executados:

  - *Feature Selection* – Seleção de tokens que representam características do documento.
  - *Feature Weighting* – Ponderação de características com atribuição de peso aos tokens selecionados.

- *Feature Projection* – Projeção de características em modelo de menor dimensão.
- *Feature Scaling* – Transformação dimensional do espaço de características.
- *Instance Selection* – Seleção de instâncias para redução do espaço amostral.

3.6.2 Treinamento do Modelo

Nesta fase, o modelo computacional escolhido passa pela indução de treinamento.

Procedimentos executados:

- Particionamento — Definição dos conjuntos de dados para treinamento, teste e avaliação.
- Balanceamento Amostral — Correções e ajustes na distribuição de amostras por conjuntos.
- Treinamento do Algoritmo — Apresentação do conjunto de treinamento ao algoritmo escolhido.
- Testagem e Tunagem — Ciclo de testes e ajustes de parâmetros e hiperparâmetros para atingir melhor desempenho possível.
- Construção do Modelo Treinado — Persistência do modelo de aprendizagem e suas informações de treinamento (o conhecimento aprendido).

3.6.3 Validação e Avaliação

Fase dedicada aos testes qualitativos do modelo treinado.

Procedimentos executados:

- Escolha de Métricas — Determinação de índices de avaliação (Acurácia, Precisão, Recall, F-Measure, etc.).
- Escolha de Método — Determinação do método de execução de testes avaliativos (*leave-one-out*, *k-fold cross-validation*, etc.).
- Validação — Execução do modelo de avaliação definido, cômputo e sumarização de resultados.
- Avaliação — Cálculo dos valores de índices de avaliação, comparação com modelo de base.

3.7 Análise Semântica e Processamento de Linguagem Natural

Segundo Santos e Merschmann (2020), a análise semântica de texto refere-se à compreensão do significado das palavras no texto, seja pela utilização de dicionários ou pela extração de seus contextos, sendo que a detecção de palavras-chave é uma das possíveis técnicas aplicáveis para esta finalidade.

O Processamento de Linguagem Natural (NLP, do inglês *Natural Language Processing*), é a área do conhecimento que estuda técnicas computacionais para representação e análise

automática da linguagem humana. Devido a própria característica da linguagem natural, NLP é amplamente empregado em análises semânticas de textos, apresentando tarefas relacionadas com este propósito (FISHER; GARNSEY; HUGHES, 2016; KUMAR; RAVI, 2016; YOUNG et al., 2018; CHENG; NAZARIAN; BOGDAN, 2020). A implementação destas técnicas frequentemente está relacionada com abordagens complexas de *Machine Learning* e *Deep Learning* (YOUNG et al., 2018).

3.8 Considerações Finais sobre o Referencial Teórico

Este capítulo apresentou conceitos fundamentais relacionados aos temas Análise Visual, Visualização da Informação e Classificação de Texto. Estabelecendo conceitos, descrevendo aplicações e fornecendo exemplos relacionados com as técnicas e métodos abordados. Os próximos capítulos detalham como estes conceitos foram aplicados neste estudo de caso.

4 Problemas de Classificação Abordados no Estudo

Este capítulo apresenta detalhadamente os dois problemas de classificação abordados durante o estudo.

4.1 O Problema da Classificação de Finalidade

A Classificação de Finalidade tem como objetivo estabelecer para que o recurso financeiro está sendo empregado.

Observando pela perspectiva laboral, a execução de uma atividade pode ser relacionada com uma ou mais finalidades. Da mesma forma se dá pela perspectiva financeira, onde um montante pode ser relacionado à várias finalidades. Contudo, pela visão da execução orçamentária e da gestão administrativa, cada lançamento financeiro contábil é a materialização da menor unidade que compõe a execução orçamentária, por isso cada registro poderá ter atribuído a si somente uma finalidade.

Um conjunto de cinco finalidades foram elencadas para serem relacionadas aos registros contábeis. Sendo três relacionadas com as atividades-fim da universidade: i) Ensino; ii) Pesquisa; e, iii) Extensão Universitária. E mais duas finalidades relativas às atividades-meio da instituição: iv) Gestão Administrativa; e, v) Fornecimento de Infraestrutura.

A Tabela 5 apresenta as classes relacionadas no conjunto, com suas respectivas quantidades de registros e representatividade percentual no *dataset*. Observa-se que a distribuição de registros entre as classes é altamente desbalanceada, sendo a categoria majoritária ("Ensino") mais de dez vezes maior do que a categoria minoritária ("Extensão Universitária").

Tabela 5 – Conjunto de classes para classificação de Finalidade.

Classe Finalidade	Quantidade de Registros	Percentual do *Dataset*
Ensino	1574	45,60%
Pesquisa	361	10,46%
Extensão Universitária	150	4,34%
Gestão Administrativa	1141	33,05%
Fornecimento de Infraestrutura	226	6,55%
Total	**3452**	**100,00%**

Fonte: Produzida pelo autor.

➤ Exemplo Ilustrativo de Classificações de Finalidade

Para exemplificar a classificação de finalidade atribuída aos lançamentos contábeis, suponhamos que um recurso financeiro é destinado à uma unidade universitária (e.g., o orçamento

geral da unidade). Então, a unidade divide este recurso e repassa partes dele aos seus centros de custo subordinados (e.g., orçamento local do centro de custo). Assim, cada centro de custo utiliza o recurso recebido para as atividades inerentes ao propósito do centro de custo (e.g., aquisição de materiais, pagamento de diárias, contratação de serviços, solicitação de itens do almoxarifado, etc.).

Considerando o exemplo do cenário exposto:

- O registro contábil de crédito, referente ao orçamento geral da unidade, pode ser classificado com a finalidade de Gestão Administrativa.
- Cada registro contábil de débito, referente aos repasses para os centros de custo subordinados da unidade, podem ser classificados com diferentes finalidades.

  — Repasse ao centro de custo Departamento de Ensino, pode ser classificado com a finalidade Ensino.

  — Repasse ao centro de custo Projeto Cultural, pode ser classificado com a finalidade Extensão Universitária.

  — Repasse ao centro de custo Recursos Humanos, pode ser classificado com a finalidade Gestão Administrativa.
- No centro de custo, o registro contábil de crédito, referente ao repasse recebido da unidade universitária, pode ser classificado como a finalidade Gestão Administrativa.

  — O registro contábil de débito, referente ao pagamento de diária para participação em banca examinadora, pode ser classificado com a finalidade Ensino.

  — O registro contábil de débito, referente ao pagamento de diária para participação no Conselho Universitário, pode ser classificado com a finalidade Gestão Administrativa.

  — O registro contábil de débito, referente à aquisição de papel sulfite para a secretaria do departamento, pode ser classificado com a finalidade Gestão Administrativa.

  — O registro contábil de débito, referente à aquisição de papel sulfite para uso dos docentes do departamento, pode ser classificado com a finalidade Ensino.

  — O registro contábil de débito, referente ao pagamento de taxa de inscrição em evento científico, pode ser classificado com a finalidade Pesquisa.

Nota-se neste exemplo que um mesmo montante financeiro pode gerar vários lançamentos contábeis, de modo que cada registro possui sua devida classificação de finalidade relacionada. Constata-se então, que este é um problema de classificação multiclasse e monorrótulo.

4.2 O Problema da Classificação de Categoria

A Classificação de Categoria tem como objetivo estabelecer em que o recurso financeiro está sendo empregado.

O conjunto total de categorias foi elaborado com termos indicados por especialistas no domínio da aplicação, membros do grupo de trabalho relacionados com a área de gestão

administrativa. Além dos termos que compõem os nomes das classes de categorias, este grupo também foi responsável pela definição do conjunto de palavras-chave que estão relacionadas com cada categoria. Sendo que cada categoria filha herda as palavras-chave de seu pai, possuindo adicionalmente um conjunto próprio de palavras-chave para distinguir-se de suas categorias irmãs.

Descartando-se lançamentos com valores nulos, observa-se que o registro contábil pode ser essencialmente de dois tipos, crédito ou débito. Consequentemente, é possível afirmar que nenhum registro do conjunto de dados poderá ter classificação nula. Pois em casos extremos, onde não seria possível especificar uma categoria discriminativa, ainda seria possível atribuir a categoria raiz de acordo com o tipo de lançamento contábil, ou atribuir-se uma categoria mais genérica se for possível identificar o ramo a qual o lançamento pertence. Desta forma, foram estabelecidas 134 categorias organizadas hierarquicamente em duas árvores distintas, tendo cada árvore o seu tipo de registro como categoria raíz.

A Tabela 6 apresenta as 13 categorias relacionadas com os registros do tipo crédito. Ao passo que a Tabela 7 resume as 121 categorias disponíveis para os registros do tipo débito. A análise da representatividade das classes de categorias apresentadas nestas tabelas, demonstra um expressivo desbalanceamento no conjunto de dados experimental, com categorias sendo centenas de vezes maiores que outras.

Tabela 6 – Conjunto de classes para classificação de categorias de crédito.

Identificação Hierárquica	Descrição da categoria de crédito	Registros e percentual no *dataset*
receitas	Receitas	1 (0,03%)
receitas.**convenio**	Convênios institucionais	9 (0,26%)
receitas.convenio.**fomento**	Agência de fomento	10 (0,29%)
receitas.convenio.**governo**	Convênio com governos	3 (0,09%)
receitas.convenio.**privado**	Iniciativa privada	7 (0,20%)
receitas.**custeio**	Recursos de custeio	80 (2,32%)
receitas.custeio.**orcamentario**	Dotação orçamentária de custeio	73 (2,11%)
receitas.custeio.**repasse**	Repasse de recursos	20 (0,58%)
receitas.custeio.**ressarcimento**	Ressarcimento de recursos	484 (14,02%)
receitas.**propria**	Receita própria	224 (6,49%)
receitas.propria.**concurso**	Inscrição em concurso	44 (1,27%)
receitas.propria.**curso**	Inscrição em curso	105 (3,04%)
	Total categorias de crédito	**1060 (30,71%)**

Fonte: Produzida pelo autor.

As colunas *"Identificação Hierárquica"* das Tabelas 6, 7, 8 e 9 provêm informação da relação de generalização/especialização entre as classes de categorias. As últimas palavras destacadas em negrito nesta coluna, referem-se ao termo escolhido como palavra identificadora da categoria, um termo curto e apropriado para ser utilizado como legenda em ferramentas gráficas de visualização da informação.

Tabela 7 – Conjunto de classes para classificação de Categorias de Débito.

Identificação Hierárquica	Descrição da Categoria de Débito	Registros
despesas.**adiantamento**	Aplicação em conta de adiantamento	34 (0,98%)
despesas.anuidade.**associacao**	Anuidade de associações	2 (0,06%)
despesas.**aquisicoes**	Aquisições e contratações	3 (0,09%)
despesas.aquisicoes.**material**	Aquisição de materiais	8 (0,23%)
despesas.aquisicoes.material.**[...]**	Detalhado na Tabela 8	500 (14,48%)
despesas.aquisicoes.**servico**	Contratação de serviços	3 (0,09%)
despesas.aquisicoes.servico.**[...]**	Detalhado na Tabela 9	253 (7,33%)
despesas.**correio**	Correios	11 (0,32%)
despesas.**devolucao**	Devolução de recursos	28 (0,81%)
despesas.diarias.**internacional**	Diárias internacionais	8 (0,23%)
despesas.diarias.**nacional**	Diárias no Brasil	302 (8,75%)
despesas.estagiarios.**diretos**	Estagiários diretamente contratados	7 (0,20%)
despesas.estagiarios.**docencia**	Estágio supervisionado em docência	112 (3,24%)
despesas.estagiarios.**terceirizados**	Estagiários contratados por instituição credenciada	6 (0,17%)
despesas.**estoque**	Solicitação de itens do estoque	160 (4,64%)
despesas.**estorno**	Estorno de recursos	28 (0,81%)
despesas.financeiro.**colaborador**	Ajuda de custo a colaboradores eventuais	76 (2,20%)
despesas.financeiro.**estudante**	Auxílio financeiro a estudantes	209 (6,05%)
despesas.financeiro.**pesquisador**	Auxílio financeiro a pesquisadores	47 (1,36%)
despesas.**pessoal**	Despesas com pessoal	59 (1,71%)
despesas.pessoal.**13salario**	13º Salário dos servidores	19 (0,55%)
despesas.pessoal.**beneficios**	Benefícios dos servidores	14 (0,41%)
despesas.pessoal.beneficios.**auxilios**	Auxílios aos servidores	6 (0,17%)
despesas.pessoal.beneficios.**transporte**	Vale transporte	13 (0,38%)
despesas.pessoal.**ferias**	Férias dos servidores	31 (0,90%)
despesas.pessoal.**formacao**	Auxílio formação e qualificação	4 (0,12%)
despesas.pessoal.**previdencia**	Previdência dos servidores	61 (1,77%)
despesas.pessoal.**vencimentos**	Vencimentos dos servidores	43 (1,25%)
despesas.**taxas**	Taxas	10 (0,29%)
despesas.taxas.**inscricao**	Taxa de inscrição e publicação em eventos	17 (0,49%)
despesas.**transferencias**	Transferência de recursos	112 (3,24%)
despesas.transportes.**aereo**	Passagens aéreas	16 (0,46%)
despesas.transportes.**combustivel**	Combustível	4 (0,12%)
despesas.transportes.**detin**	Detin	117 (3,39%)
despesas.transportes.**rateio**	Rateio de viagem	5 (0,14%)
despesas.transportes.**ressarcimento**	Ressarcimento de despesas com transportes	15 (0,43%)
despesas.transportes.**terrestre**	Passagens terrestres	49 (1,42%)
	Total categorias de débito	**2392 (69,29%)**

Fonte: Produzida pelo autor.

Os subconjuntos de categorias relacionadas com a aquisição de materiais ou contratação de serviços, destacados respectivamente nas Tabelas 8 e 9, foram criados de forma automatizada utilizando-se dados do sistema da Bolsa Eletrônica de Compras do Estado de São Paulo (BEC) (BEC, 2022a; BEC, 2022b). Os nomes das classes de produtos deste sistema, foram condensados em uma única palavra e utilizados como categorias neste problema de classificação. Sendo que para cada classe disponível no sistema, os nomes completos dos grupos de produtos vinculados à classe, foram utilizados para formar o conjunto específico de palavras-chave da categoria.

Tabela 8 – Subcategorias de Aquisição de Materiais.

Identificação Hierárquica	Descrição da Subcategoria de Aquisição de Materiais	Registros
...material.**acessibilidade**	Equipamentos,Maquinas E Artigos Para Acessibilidade Das Pessoas Com Ne (BEC:886)	3 (0,09%)
...material.**aeronautico**	Aeronaves, Acessorios, Equipamentos E Componentes (BEC:836)	1 (0,03%)
...material.**alarmes**	Sistemas De Alarme,Sinalizacao, Deteccao Para Seguranca E Equipament (BEC:869)	1 (0,03%)
...material.**alimenticio**	Generos Alimenticios (BEC:893)	50 (1,45%)
...material.**amarracao**	Cordas, Cabos E Correntes (BEC:850)	2 (0,06%)
...material.**audiovisual**	Equipamentos Fotograficos, Filmograficos E Fonograficos (BEC:873)	12 (0,35%)
...material.**autopecas**	Pecas E Acessorios Para Automoveis, Motocicletas, Ciclomoto-Res, Moton (BEC:843)	11 (0,32%)
...material.**bombeamento**	Bombas E Compressores (BEC:853)	1 (0,03%)
...material.**comunicacao**	Equipamentos De Comunicacao, Deteccao E Radiacao (BEC:864)	2 (0,06%)
...material.**construcao**	Materiais Para Construcao E Pavimentacao (BEC:863)	13 (0,38%)
...material.**decorativo**	Artigos, Utensilios E Utilidades De Uso Geral (BEC:876)	15 (0,43%)
...material.**desportivo**	Equipamentos E Materiais Para Recreacao E Desporto (BEC:882)	2 (0,06%)
...material.**eletroeletronicos**	Componentes De Equipamentos Eletricos E Eletronicos (BEC:865)	24 (0,70%)
...material.**encanamento**	Canos, Tubos, Mangueiras E Acessorios (BEC:857)	7 (0,20%)
...material.**energia**	Condutores Eletricos E Equipamentos De Forca E Distribuicao (BEC:867)	40 (1,16%)
...material.**escritorio**	Artigos E Utensilios De Escritorios (BEC:879)	51 (1,48%)
...material.**ferragens**	Ferragens E Abrasivos (BEC:860)	18 (0,52%)
...material.**ferramentas**	Ferramentas Manuais (BEC:859)	22 (0,64%)
...material.**filtragem**	Equipamentos Para Purificacao E Filtragem De Agua (BEC:856)	1 (0,03%)
...material.**fluidos**	Combustiveis, Oleos, Lubrificantes E Ceras (BEC:895)	5 (0,14%)
...material.**gases**	Instrumentos E Equipamentos De Controle De Medicao E De Gases Comprimi (BEC:872)	5 (0,14%)
...material.**hidraulica**	Equipamentos De Instalacoes Hidraulicas, Sanitarias E De Ca-Lefacao (BEC:855)	6 (0,17%)
...material.**higiene**	Artigos De Higiene (BEC:889)	6 (0,17%)
...material.**hospitalar**	Equipamentos E Artigos De Uso Medico, Odontologico E Hospitalar (BEC:871)	24 (0,70%)
...material.**iluminacao**	Lampadas Para Iluminacao De Ambiente E Aparelhos De Iluminacao (BEC:868)	6 (0,17%)
...material.**industrial**	Maquinas E Equipamentos Para Industrias Especializadas (BEC:846)	2 (0,06%)
...material.**informatica**	Informatica (BEC:890)	60 (1,74%)
...material.**limpeza**	Equipamentos E Materiais Para Limpeza, Dedetizacao E Esteri-Lizacao De (BEC:883)	12 (0,35%)
...material.**madeira**	Tabuas, Compensados De Madeira, Esquadrias E Portas De Madeira, Ferro (BEC:862)	5 (0,14%)
...material.**manufaturados**	Materiais Manufaturados Nao Metalicos (BEC:897)	2 (0,06%)
...material.**manuseio**	Maquinas E Equipamentos Para Manuseio De Material (BEC:849)	2 (0,06%)
...material.**metalicos**	Barras,Chapas E Perfilados Metalicos (BEC:899)	9 (0,26%)
...material.**metroviario**	Materiais E Equipamentos Metroviarios E Ferroviarios (BEC:894)	1 (0,03%)
...material.**mobiliario**	Mobiliarios Em Geral (BEC:875)	3 (0,09%)
...material.**musical**	Instrumentos Musicais, Obras De Arte E Artesanatos (BEC:881)	2 (0,06%)
...material.**oficina**	Maquinas E Equipamentos De Oficinas De Manutencao (BEC:844)	10 (0,29%)
...material.**pintura**	Pinceis, Tintas, Vedantes E Adesivos (BEC:884)	5 (0,14%)
...material.**pneus**	Pneus E Camaras (BEC:842)	6 (0,17%)
...material.**radiotv**	Equipamentos E Componentes Para Emissoras De Radio E Televisao (BEC:866)	2 (0,06%)
...material.**recipientes**	Recipientes E Materiais Para Acondicionamento E Embalagem (BEC:885)	2 (0,06%)
...material.**refeitorio**	Equipamentos E Utensilios Para Refeitorio, Copa E Cozinha (BEC:877)	25 (0,72%)
...material.**refrigeracao**	Equipamentos Para Refrigeracao, Condicionamento E Purifica Cao De Ar (BEC:851)	9 (0,26%)
...material.**sinalizacao**	Placas E Acessorios De Identificacao E Sinalizacao (BEC:901)	1 (0,03%)
...material.**textil**	Tecidos, Couros, Peles, Aviamentos, Barracas E Bandeiras (BEC:887)	6 (0,17%)
...material.**utilidades**	Equipamentos,Maquinas E Materiais Para Servicos Gerais (BEC:845)	1 (0,03%)
...material.**valvulas**	Valvulas (BEC:858)	2 (0,06%)
...material.**vestuario**	Vestuarios, Equipamentos Individuais E Insignias (BEC:888)	5 (0,14%)
	Total categorias de relacionadas com aquisição de materiais	500 (14,48%)

Fonte: Produzida pelo autor.

Tabela 9 – Subcategorias de Contratação de Serviços.

Identificação Hierárquica	Descrição da Subcategoria de Contratação de Serviços	Registros
...servico.**especializado**	Servicos Especializados (BEC:2)	134 (3,88%)
...servico.**instalacao**	Servicos De Instalacoes/Montagens (BEC:4)	5 (0,14%)
...servico.**locacao**	Servicos De Locacoes (BEC:7)	1 (0,03%)
...servico.**locomocao**	Servicos De Transportes, Manuseios De Materiais, Acondicionamentos E Armazenagens (BEC:6)	1 (0,03%)
...servico.**manutencao**	Servicos De Manutencoes/Conservacoes De Bens Moveis (BEC:5)	57 (1,65%)
...servico.**reforma**	Servicos De Adaptacoes, Reparos, Reformas E Instalacoes Em Obras Civis, De Engenharia E De Construcoes (BEC:3)	5 (0,14%)
...servico.**seguro**	Seguros	2 (0,06%)
...servico.**servicosgerais**	Servicos Gerais (BEC:8)	48 (1,39%)
	Total categorias de relacionadas com contratação de serviços	253 (7,33%)

Fonte: Produzida pelo autor.

A Tabela 10 apresenta 32 categorias não utilizadas explicitamente na classificação do conjunto de dados experimental. Sendo 31 categorias do tipo débito e uma categoria de crédito. A análise das categorias não utilizadas no conjunto de dados experimental, demonstrou que a maioria destas categorias estão relacionadas com a aquisição de materiais esporádicos ou são categorias que não fazem parte do escopo da unidade universitária.

Tabela 10 – Categorias não utilizadas no conjunto de dados experimental.

Motivo de Ausência	Categoria
Compra esporádica	Animais Vivos (BEC:892) Embarcacoes - Acessorios, Equipamentos E Componentes (BEC:837) Equipamentos E Artigos De Uso Veterinario (BEC:870) Equipamentos Para Construcao, Conservacao De Rodovias,Mineracao E Esca (BEC:848) Equipamentos, Materiais E Acessorios Para Combate A Incendio,Resgate E (BEC:852) Estruturas E Andaimes Pre-Fabricados (BEC:861) Ferrovias - Acessorios, Equipamentos E Componentes (BEC:839) Fornos, Caldeiras E Reatores (BEC:854) Maquinas E Equipamentos Agricolas E Para Pecuaria (BEC:847) Maquinas E Equipamentos Para Escritorios (BEC:878) Maquinas E Equipamentos Para Fins Didaticos (BEC:838) Material Belico (BEC:834) Repasse de recursos Servicos De Estudos, Pesquisas E Projetos (BEC:1) Suprimentos Agricolas (BEC:891) Tratores (BEC:841) Veiculos Rodoviarios (BEC:840)
Categoria raiz ou intermediária	Anuidades Auxílios financeiros Despesas Despesas com estagiários Despesas com transportes Diárias Rateio de despesas
Responsabilidade de outrem	Anuidade de revistas e periódicos Despesas de ações trabalhistas Livros, Mapas E Outras Publicacoes (BEC:880) Servicos Publicos Terceirizados (BEC:9) Vale alimentação
Não ocorreu	Anuidade de conselho de classe Inscrição em evento (Crédito)
Não se aplica	Pedágio

Fonte: Produzida pelo autor.

Das 102 categorias com registros no conjunto de dados experimental, temos 12 categorias de crédito que foram utilizadas na classificação de 1060 registros, representando 30.71% do conjunto de dados. E 90 categorias de débito que foram utilizadas na classificação de 2392 registros, representando 69.29% do conjunto de dados.

Considerando as informações apresentadas, nota-se que a classificação de categorias dos registros contábeis, pode ser encarado como um problema de classificação hierárquica, monorrótulo e com predição opcional em nós folha.

4.3 Considerações Finais sobre os Problemas de Classificação Abordados

Este capítulo apresentou os dois problemas de classificação abordados neste trabalho, demonstrando detalhes dos problemas e informando as diretrizes estabelecidas para cada problema abordado. O próximo capítulo abordará o algoritmo de classificação estabelecido para lidar com os problemas apresentados.

5 A Solução de Classificação de Texto

A solução para automatização da classificação dos registros financeiros e contábeis, foi o desenvolvimento de um método de classificação e integração deste ao sistema institucional. Este método de classificação foi idealizado para ser um classificador textual simples e genérico, que possa ser utilizado em diferentes problemas de classificação de texto.

Este capítulo aborda o método de classificação desenvolvido e sua possível integração ao sistema SisADM.

5.1 O Método de Classificação

O método de classificação proposto utiliza a abordagem lexical para análise de texto, que denota a criação de dicionários para representação da seleção de características de interesse nos documentos de texto. Neste sentido, o conjunto de palavras-chave atribuídos às classes, é a implementação do dicionário léxico que representa semanticamente as características das classes envolvidas no processo de classificação. E o modelo de representação vetorial *Bag of Words* (BoW) é utilizado para construção da matriz de termos do documento de texto, matriz que é confrontada pelos procedimentos comparativos, com o dicionário de palavras-chave da classe observada. Ou seja, o método desenvolvido é um algoritmo que utiliza a abordagem lexical, implementada pela contagem de ponderada da ocorrência de palavras-chave.

O Algoritmo 1 descreve este funcionamento geral do método de classificação. Sendo este, um algoritmo de com comportamento preguiçoso (do inglês, *lazy*), pois não realiza nenhuma aprendizagem de parâmetros ou função discriminativa. Todo conhecimento necessário está incutido nos dicionários de palavras-chave presentes em cada classe apresentada ao classificador.

O processo de classificação começa com a recepção do documento de texto e do conjunto de classes elegíveis à serem atribuídas ao documento. De modo geral, o classificador realiza comparação de palavras e expressões utilizadas em documentos, com o rol de palavras-chave relacionadas às classes elegíveis para classificação. O cômputo da similaridade ponderada entre os termos analisados é utilizado para atribuição de pontuação às classes, então o classificador utiliza a abordagem *One-vs-All* para estabelecer o ranqueamento das classes elegíveis.

Inicialmente todas as classes do conjunto recebem pontuação nula (zero), então, para cada classe do conjunto recebido como parâmetro de entrada, o classificador realiza até cinco procedimentos comparativos entre o documento e a classe analisada. Os procedimentos comparativos podem incrementar a pontuação da classe analisada, utilizando valores e critérios especificados nos parâmetros de classificação.

Ao final do processamento o classificador realiza o ranqueamento das classes elencadas e apresenta duas saídas que podem atender a diferentes problemas de classificação: i) classe atribuída ao documento de texto, sendo esta a principal saída do classificador, ou ii) *ranking* das classes analisadas, uma saída complementar que pode ser interpretada por ferramentas externas.

Basicamente, problemas de classificação monorrótulo são atendidos pela saída direta da classe atribuída ao documento. Ao passo que, problemas de classificação multirrótulo ou problemas de classificação hierárquica, podem ser resolvidos com a interpretação da saída do ranqueamento das classes analisadas.

Algoritmo 1: Classificador de Texto Baseado em Palavras-Chave (*Keywords*)

ENTRADAS : Documento, Classes
SAÍDAS : ClasseAtribuida, ScoreRanking

```
1  início
2      /* Inicialização */
3      para cada classe c de Classes faça
4          ranqueamento [c] ← Score(0);
5      fim
6      /* Procedimentos Comparativos por Palavras-Chave */
7      se não Vazio(Documento.conteudo) então
8          texto ← getPalavrasNormalizadas(Documento.conteudo);
9          para cada classe c de Classes faça
10             para cada keyword em c.keywords faça
11                 se texto.tamanho éSemelhante keyword.tamanho então
12                     comparacaoDeAltaRelevancia(texto, keyword);
13                 fim
14                 destinar kw para simples[] ou compostas[];
15             fim
16             compararPorPalavrasChaveCompostas(texto, compostas[]);
17             compararPorPalavrasChaveSimples(texto, simples[]);
18         fim
19     fim
20     /* Procedimentos Comparativos por Nome e Título */
21     se não Vazio(Documento.titulo) então
22         compararPorNomeParcialDeClassse(Documento.titulo);
23         compararPorNomeCompletoDeClassse(Documento.titulo);
24     fim
25     /* Aplicando One-vs-All */
26     construa ScoreRaking com apenas ranqueamentos válidos;
27     ordene ScoreRaking em ordem decrescente;
28     /* Determinar a Classe Atribuída */
29     se ScoreRaking tem um líder único então
30         ClasseAtribuida ← ScoreRaking[0];
31     fim
32 fim
```

O método proposto baseia-se no conceito de similaridade de expressões textuais, uma

das formas de correspondência aproximada de strings apresentadas por Hall e Dowling (1980).

5.2 Cálculo da Similaridade Ponderada

Primeiro, é necessário estabelecer se as palavras comparadas possuem tamanhos semelhantes, para isto estabeleceu-se que a variação máxima aceita na diferença de tamanhos de palavras é de ±50%, valor atribuído com base na observação das possíveis flexões das palavras-chave identificadas nas classes conhecidas. Desta forma, palavras como "Departamento" e "Depto", são consideradas incomparáveis por não terem tamanhos semelhantes. O nível de similaridade considerado para os casos de palavras de tamanhos muito diferentes é zero.

Para duas palavras de tamanhos semelhantes, o cálculo de similaridade entre as palavras é realizado com o algoritmo *Longest Common Subsequence* (LCS) proposto por Hirschberg (1977). A escolha deste algoritmo foi motivada por esta ser uma boa técnica para identificação de semelhanças entre duas expressões textuais curtas, podendo também ser utilizada com mais expressões (IRVING; FRASER, 1992). Utilizado tradicionalmente na comparação de duas palavras, LCS implicitamente abstrai flexões de palavras e erros de digitação, pois considera a sequência de caracteres que são comuns para às palavras observadas (PIKIES; ALI, 2019; PRADHAN; GYANCHANDANI; WADHVANI, 2015; WANG; DONG, 2020; PRAKOSO; ABDI; AMRIT, 2021). O valor de similaridade é calculado pela razão entre o tamanho do termo gerado pela LCS e o tamanho da maior palavra utilizada.

$$SIMILARIDADE = \frac{comprimento\_lcs}{comprimento\_maior\_palavra}$$

Uma subsequência com três caracteres, no caso da maior palavra analisada possuir quatro letras, representa um valor de similaridade de 3/4 (75%). Este mesmo valor de similaridade pode ser observado em relações como 6/8 ou 9/12. Contudo, ao consideramos as probabilidades da análise combinatória para os arranjos de caracteres nas relações 6/8 ou 9/12, nota-se que o aumento das letras consideradas na relação, incute maior relação semântica para o conceito de similaridade, visto que a probabilidade de obter uma correspondência de nove caracteres em um conjunto de tamanho doze é bem mais difícil do que obter a correspondência de três em quatro. Desta forma, optou-se por ponderar o valor de similaridade utilizando a diferença Δ dos tamanhos das palavras analisadas.

$$\Delta = |comprimento\_palavra1 - comprimento\_palavra2|$$

$$SIMILARIDADE\_PONDERADA = SIMILARIDADE \times \left(1 + \left(\frac{\Delta}{100}\right)\right) \tag{5.1}$$

A Equação 5.1 apresenta o cálculo do valor de similaridade ponderada. O procedimento de ponderação incrementa em 1% o valor da similaridade computada pela LCS, para cada caractere de diferença na relação entre as palavras analisadas. Por exemplo, a relação 3/4 apresenta 1 caractere de diferença, então 75.00% × 1.01 resulta no valor ponderado de 75.75%. Da mesma

forma, observa-se que a relação 7/9 possui 2 caracteres de diferença, logo considera-se o valor ponderado de 79.33%, que é obtido pelo cálculo 77.77% × 1.02

A Tabela 11 apresenta exemplos de palavras comparadas, com a demonstração da maior subsequência encontrada entre elas. O cálculo de valor de similaridade pode ser observado nas colunas *"Grau Similaridade"* e *"% Ponderado"*, que apresentam os valores da relação entre o tamanho da LCS e o tamanho da maior das palavras analisadas.

Foram estabelecidos nove níveis de similaridade para qualificar a aproximação das palavras observadas, sendo o nível de similaridade atribuído à comparação definido pelo seu valor de referência. A coluna *"Nível de Similaridade"* da Tabela 11 apresenta qual foi o nível de similaridade atribuído à relação, destacando entre parênteses o valor de referência para atribuição de tal nível.

Tabela 11 – Similaridade de Palavras-Chave por LCS.

Palavra 1		Palavra 2		LCS	Grau Similaridade	% Ponderado	Nível de Similaridade
departamento	+	depto	=	*Incomparável*	*Não aplicável*	0,00%	Nenhuma (% = 0)
detin	+	bolsas	=		0/6 (0,00%)	0,00%	Nenhuma (% = 0)
bolsas	+	auxilio	=	o	1/7 (14,28%)	15,14%	Mínima (% >= 11)
higiene	+	limpeza	=	ie	2/7 (28,57%)	30,00%	Baixa (% >= 23)
impressora	+	imprimir	=	imprr	5/10 (50,00%)	52,50%	Média (% >= 47)
hospedar	+	hospitalar	=	hospar	6/10 (60,00%)	62,40%	Moderada (% >= 61)
casa	+	cada	=	caa	3/4 (75,00%)	75,75%	Alta (% >= 71)
auxilio	+	auxiliar	=	auxili	6/8 (75,00%)	76,50%	Alta (% >= 71)
alimenticios	+	alimentos	=	alimentos	9/12 (75,00%)	77,25%	Alta (% >= 71)
eventual	+	eventuais	=	eventua	7/9 (77,77%)	79,33%	Muito alta (% >= 79)
adequacao	+	graduacao	=	aduacao	7/9 (77,77%)	79,33%	Muito alta (% >= 79)
administracao	+	administrativo	=	administrao	11/14 (78,57%)	80,93%	Muito alta (% >= 79)
construcoes	+	construcao	=	construco	9/11 (81,81%)	83,45%	Muito alta (% >= 79)
colaborador	+	colaboradores	=	colaborador	11/13 (84,61%)	86,31%	Muito alta (% >= 79)
doutorando	+	doutorado	=	doutorado	9/10 (90,00%)	90,90%	Quase total (% >= 89)
laboratorio	+	laboratorio	=	laboratorio	11/11 (100,00%)	100,00%	Total (% = 100)

Fonte: Produzida pelo autor.

5.3 Procedimentos Comparativos

Os documentos analisados podem ter os atributos título e conteúdo. Enquanto as classes elegíveis para classificação possuem três atributos: i) identificação; ii) nome; e, iii) conjunto de palavras-chave. As palavras-chave são formadas por expressões textuais de tamanho *n-gram*, podendo ser divididas em grupos de palavras-chave compostas ou simples, havendo ou não espaços para separar as palavras da expressão.

O termo *score* designa a pontuação atribuída à classe, este é o principal valor utilizado para o ranqueamento do conjunto de classes elencadas na classificação do documento. Cada procedimento comparativo incrementa o *score* da classe, conforme registra ocorrência de similaridade entre os objetos de comparação. Os três primeiros procedimentos comparativos utilizam o conteúdo do documento e conjunto de palavras-chave. Os dois últimos estão relacionados com o nome da classe e título do documento.

5.3.1 Procedimento Comparativo I – comparacaoDeAltaRelevancia

Este é o primeiro dos procedimentos comparativos a lidar com o conjunto de palavras-chave e o conteúdo do documento. Sendo realizado somente quando o conteúdo do documento possui tamanho semelhante ao da palavra-chave analisada. Se houver o registro de similaridade *"Quase Total"* entre todo o conteúdo do documento e a palavra-chave de tamanho semelhante, serão adicionados cinco pontos no *score* da classe.

5.3.2 Procedimento Comparativo II – compararPorPalavrasChaveCompostas

Palavras-chave compostas são expressões de tamanho *n-gram*, onde há caractere de espaço separando as palavras da expressão. Se houver ocorrência da palavra-chave composta no conteúdo do documento, então adiciona-se dois pontos no *score* da classe. Esta verificação é um procedimento comparativo simples que não utiliza o conceito de similaridade.

Além do procedimento simples de verificação de ocorrência da palavra-chave composta, os termos que compõe a palavra-chave são decompostos em termos *unigram* (1-*gram*), aproveitando-se aqueles que possuem tamanho mínimo relevante para serem utilizados no procedimento de comparação por palavras-chave simples.

5.3.3 Procedimento Comparativo III – compararPorPalavrasChaveSimples

Para cada palavra-chave presente na classe analisada, verifica-se a ocorrência de similaridade com as palavras do conteúdo do documento de texto. Havendo similaridade ponderada com nível de similaridade *"Muito Alta"* ou superior, adiciona-se um ponto ao *score* da classe.

5.3.4 Procedimento Comparativo IV – compararPorNomeParcialDeClassse

Quando o documento observado possui o atributo título, este é decomposto em termos *unigram* e utilizado como conjunto de palavras-chave com maior peso. Consideram-se válidos como palavra-chave apenas os termos que possuírem tamanho mínimo relevante. Para cada ocorrência de similaridade ponderada, considerando no mínimo o nível de similaridade *"Muito Alta"*, na comparação entre os termos deste novo conjunto de palavras-chave e o conteúdo do documento, adiciona-se três pontos no *score* da classe.

5.3.5 Procedimento Comparativo V – compararPorNomeCompletoDeClassse

Outro procedimento realizado quando o título do documento está presente, é a comparação entre o nome da classe observada e o título do documento analisado. Registrando-se ocorrência de similaridade ponderada entre eles, o *score* da classe é aumentado em oito pontos, desde que o nível de similaridade seja no mínimo *"Quase Total"*.

5.4 Parâmetros de Classificação e Ranking

Os procedimentos comparativos definem valores de atributos qualitativos da classificação de classes. Sendo estes atributos: i) o *score*, que define a pontuação geral da classe; ii) a quantidade

de ocorrências registradas, que contabiliza quantas vezes um procedimento de comparação registrou reposta afirmativa; e, iii) o percentual acumulado de similaridade, que registra a soma do valor percentual de todas as ocorrências positivas de similaridade.

Os pesos atribuídos aos procedimentos de classificação foram determinados com a realização de exaustivos experimentos para o ajuste fino do classificador. Além dos parâmetros de ponderação dos procedimentos comparativos, o classificador possui mais seis parâmetros de personalização:

- Tamanho do Ranking — Estabelece a quantidade máxima de classes que poderão compor a saída de classes ranqueadas.
- Pontuação Mínima — Refere-se ao *score* mínimo necessário para que a classe observada seja incluída no ranking de classificação.
- Critérios de Desempate — Define a utilização de critérios de desempate para atribuir uma classe ao documento.
- Forçar a Classificação — Parâmetro que consiste na utilização conjunta de valor nulo para pontuação mínima e ativação da utilização de critérios de desempate.
- Normalizar Strings — Configuração para efetuar a remoção de acentuação, remoção de pontuação, remoção de espaços em branco, remoção de caracteres especiais ASCII e remoção de caracteres não-ASCII.
- Somente Minúsculas — Define que todo texto observado deverá ser transformado em letras minúsculas antes das comparações.

O primeiro valor considerado para o ranqueamento é o *score* da classe. Em caso de empate no *score*, pode-se analisar a quantidade ocorrências registradas nos procedimentos de comparação e o percentual acumulado de similaridade.

Para estabelecer o ranqueamento final, todos os atributos qualitativos são ordenados de forma decrescente, de modo que a primeira classe do ranking seja aquela com os maiores valores atribuídos. Em caso de empate total nos atributos qualitativos, o atributo de identificação da classe é utilizado como valor de ordenação no ranking, para este último caso a ordenação é crescente.

5.5 Saídas do Classificador

Com objetivo de desenvolver um classificador textual mais genérico possível, estabeleceu-se que dois tipos de saídas seriam geradas pelo algoritmo de classificação.

A saída direta (*"ClasseAtribuida"*), consiste na classe única atribuída ao documento analisado. Esta saída nem sempre estará com valor atribuído, pois a saída de classe única depende dos parâmetros estabelecidos para classificação. Se nenhuma classe elegível atingir a pontuação

mínima para ser considerada válida ou, caso o classificador não esteja habilitado para uso de critérios de desempate, não será possível a atribuição de uma classe ao documento.

A saída interpretável (*"ScoreRanking"*), é formada pelo ranqueamento ordenado das classes observadas, considerando somente aquelas que atingiram a pontuação mínima estabelecida. Da mesma forma que a saída direta, o parâmetro de corte pela pontuação mínima pode afetar esta saída, pois caso nenhuma classe seja ranqueada, será gerado um conjunto vazio como resposta.

5.6 Implementação do Classificador de Texto

O Sistema de Gestão Administrativa (SisADM) em uso na Universidade foi desenvolvido em linguagem de programação Java, utilizando-se a arquitetura J2EE (*Java 2 Platform, Enterprise Edition*). Por este motivo, optou-se pela implementação do método proposto utilizando a mesma linguagem de programação e arquitetura deste sistema.

A solução de classificação automática integrada ao SisADM foi idealizada em três módulos com atribuições distintas:

- Módulo de Integração — Responsável pela obtenção de registros do SisADM e geração do documento de texto a ser classificado. Também responsável pela definição dos parâmetros de entrada do classificador e por realizar a interpretação da saída do classificador.
- Módulo de Similaridade — Responsável pelo cômputo de similaridade textual.
- Módulo de Classificação — Responsável pelo processamento do documento de texto e realização do processo de classificação, conforme parâmetros de entrada.

5.7 Integração do Classificador de Texto ao SisADM

A solução de classificação de texto abordada neste capítulo, consiste na integração do método de classificação ao sistema institucional.

5.7.1 Parametrização da Entrada do Classificador

O módulo de integração utiliza informações disponíveis no registro financeiro e contábil, para realizar a parametrização dos dados de entrada do classificador. Neste sentido, considera-se viável a utilização da informação referente ao tipo de registro contábil, para escolha da árvore de categorias elegíveis para o registro (classes de crédito ou débito).

Também considera-se viável a utilização da informação referente à classificação contábil (Classificação Econômica), quando esta for uma classificação discriminativa (e.g., 3.3.90.30.10–Gêneros Alimentícios, 3.3.90.33.42–Passagens Aéreas, etc.) ou relacionada à poucas categorias (e.g., 3.3.90.14.01–Diárias Pessoal Civil). Estas informações podem funcionar como mecanismos de atenção, indicando ao classificador quais são as classes provavelmente mais relevantes do conjunto a ser analisado.

5.7.2 Leitura das Saídas do Classificador

Quando o algoritmo classificador atribui uma classe ao documento analisado, utiliza-se a *"ClasseAtribuida"* como resposta ao SisADM. Contudo, na ausência de classe específica atribuída, poderá ser necessária a intervenção do módulo de integração. Por exemplo, para o problema de classificação de finalidade, o módulo de integração poderá definir alguma classe padrão para ser atribuída ao registro analisado.

Considerando que o problema de classificação de categoria utiliza conjunto de classes que se organizam de forma hierárquica. Considerando também, que o classificador desenvolvido não possui conhecimento da hierarquia das classes envolvidas. Torna-se plausível a interpretação do *"ScoreRanking"* para realização de desempate por generalização. Desta forma, caso exista empate entre as primeiras categorias do *ranking* e caso estas categorias sejam irmãs, pode-se considerar a atribuição da categoria pai para o registro observado.

5.8 Considerações Finais sobre a Solução de Classificação de Texto

Este capítulo demonstrou o método de classificação desenvolvido e possibilidades de integração deste método ao sistema institucional legado. O Capítulo 6 demostrará estudo avaliativo do método de classificação.

Ressalta-se que na implantação da solução de classificação de texto, optou-se por utilizar conjunto de classes separados pelos tipos (despesas e receitas) na realização de classificação de categoria, pois o número de classes possíveis é elevado e esta separação melhora a performance da execução do procedimento de classificação. E quanto a classificação de finalidade, optou-se por não utilizar nenhuma finalidade como padrão, mas algumas informações de funcional programática foram utilizadas para direcionar a classificação à finalidades específicas. Por exemplo, o projeto-atividade código 12.392.1043.5306 implica na finalidade Extensão Universitária.

6 Avaliação do Método de Classificação

Este capítulo apresenta o estudo empírico realizado para a avaliação do método de classificação proposto no Capítulo 5.

Os dois problemas de classificação descritos no Capítulo 4 foram abordados nos experimentos realizados neste estudo avaliativo. Primeiro, o problema de classificação de finalidade (detalhado em Seção 4.1), um problema de classificação multiclasse e monorrótulo, composto por cinco classes. O segundo problema abordado refere-se à categoria (detalhado em Seção 4.2), um problema de classificação hierárquica que possui 134 classes, sendo 122 classes destinadas a registros do tipo débito e 12 classes para registros de crédito.

As mesmas informações dos registros foram utilizadas para realização das classificações nos dois problemas abordados. Tal como o mesmo conjunto de informações foi apresentado ao método de classificação proposto e aos algoritmos de referência utilizados para comparação de desempenho.

6.1 *Dataset* Experimental

Inspirado em Zayas et al. (2017), Schröder (2018), Tao, Cui e Wenjun (2018) e Ma et al. (2022), optou-se por realizar a construção de um *dataset* experimental para testes e avaliação de possíveis classificadores. Este conjunto de dados experimental foi criado com base nos registros financeiros e contábeis do sistema SisADM, considerando somente os dados da Faculdade de Engenharia – Campus de Bauru (FEB) relativos ao ano de 2019. Ano escolhido por este ser o de execução orçamentária mais recente considerando o período pré-pandemia (UNA-SUS, 2020). Desta forma, espera-se que o conjunto de dados represente condições normais de execução orçamentária, além de refletir a maturidade de implantação do sistema SisADM na Faculdade.

O conjunto de dados (*dataset*) utilizado para realização dos testes de classificação, foi estabelecido utilizando-se classes e palavras-chave definidas por especialistas no domínio da aplicação. Cada registro foi transformado em um documento texto compatível com a entrada do classificador, sendo o título do documento idêntico ao atributo título do registro do *dataset*, e o conteúdo do documento composto pela concatenação das demais informações presentes no registro.

Os registros dispõem de informações relativas à classificação contábil do lançamento e informações relacionadas à origem ou motivação de emprego do recurso financeiro. Estas informações estão distribuídas em doze colunas de informativas, das quais somente as quatro primeiras são de preenchimento obrigatório.

Relação de colunas informativas dos registros e descrição de seu possível conteúdo:

1. tipo (obrigatório) — Tipificação básica do registro (despesa ou receita).

2. origem (obrigatório) — Identificação e tipo de procedimento ou ato administrativo que originou o registro contábil.

3. titulo (obrigatório) — Título curto que representa o registro em extratos financeiros.

4. descricao (obrigatório) — Texto descritivo completo do registro contábil no sistema legado.

5. codigoClassificacaoEconomica (opcional) — Código contábil relacionado com a natureza da despesa, determinado conforme legislação vigente.

6. descricaoClassificacaoEconomica (opcional) — Texto descritivo da natureza de despesa disposto na legislação.

7. codigoProjetoAtividade (opcional) — Código de classificação orçamentária funcional programática, determinado conforme legislação vigente.

8. descricaoProjetoAtividade (opcional) — Texto descritivo da funcional programática disposto na legislação.

9. descricaoProcesso (opcional) — Título atribuído ao processo de documentação ao qual o lançamento contábil está vinculado.

10. justificativaSolicitante. (opcional) — Texto informado pelo solicitante ao iniciar um ato administrativo que origina o registro contábil.

11. gruposProdutos (opcional) — Lista de grupos de produtos que estão envolvidos com o lançamento contábil, elaborada a partir dos registros de compras do sistema legado.

12. classesProdutos (opcional) — Lista das classes de produtos que estão envolvidas no lançamento contábil, elaborada a partir dos registros do sistema legado de compras.

6.2 Implementação e Algoritmos de Referência

Os experimentos realizados com o classificador proposto, foram realizados com uma implementação do algoritmo em linguagem de programação Java, disponibilizada no Repositório GitHub[1]. Além da implementação do algoritmo, foi realizada a implementação da aplicação de interface de integração do classificador, responsável por realizar a leitura dos *datasets*, parametrizar o classificador, submeter as informações para o classificação e processar as saídas do classificador.

Como o método de classificação proposto possui um comportamento preguiçoso, inicialmente foram selecionados três algoritmos de igual comportamento para o estabelecimento de uma base de referência:

- *K-Nearest Neighbour* (k-NN) (AHA; KIBLER; ALBERT, 1991)
- *KStar* (K^*) (CLEARY; TRIGG, 1995);
- *Locally Weighted Learning* (LWL) (FRANK; HALL; PFAHRINGER, 2003)

[1] https://github.com/camargo-we/keyword-based-classifier/

Em seguida, foram selecionados mais três algoritmos tradicionais de aprendizado de máquina, os quais são frequentemente empregados na tarefa de classificação de texto:

- *Naive Bayes* (NB) (JOHN; LANGLEY, 1995)
- *Decision Trees* (DT) (QUINLAN, 1993)
- *Support Vector Machine* (SVM) (CHANG; LIN, 2011)

Para realização dos experimentos com os seis algoritmos selecionados como base de referência, foram utilizadas as implementações em linguagem Java disponíveis na aplicação Weka Workbench (EIBE; HALL; WITTEN, 2016), sendo todos com a configuração padrão, realizando-se testes de validação cruzada (*10-fold cross-validation*) e sem a utilização de filtros de pré-processamento.

6.3 Métrica de Avaliação

A acurácia média foi a métrica escolhida para avaliação e comparação dos métodos de classificação. Segundo Raza et al. (2019), acurácia é a métrica mais comum na avaliação de performances de classificadores, ela representa a eficácia do classificador, sendo necessário calcular a acurácia média quando utilizada em problemas de classificação multiclasse. Susmaga (2004) apresenta a matriz de confusão e informa que o valor de acurácia de uma classe (C) é calculado a partir das informações da respectiva matriz da classe.

$$ACURACIA(C) = \frac{VP + VN}{VP + VN + FP + FN}$$

Onde:

- VP = Verdadeiro Positivo — Indica a quantidade de registros que pertencem à categoria e foram identificados corretamente pelo classificador nesta categoria.
- VN = Verdadeiro Negativo — Quantidade de registros que não pertencem à categoria e não tiveram esta categoria atribuída pelo classificador.
- FP = Falso Positivo — Quantidade de registros que não pertencem à categoria, mas foram erroneamente categorizados com ela pelo algoritmo de classificação.
- FN = Falso Negativo — Indica a quantidade de registros que pertencem à categoria, mas não foram identificados pelo classificador como membros desta categoria.

Sendo o valor da acurácia média calculado pela soma dos valores de acurácia de cada classe (C), dividido pela quantidade (n) de classes observadas.

$$ACURACIA\_MEDIA = \frac{\sum_{i=1}^{n} ACURACIA(C_i)}{n}$$

6.4 Parametrizações

Os parâmetros de ponderação de pesos aos procedimentos comparativos foram utilizados em seus valores padrões. Quanto aos parâmetros de personalização do classificador, para o problema de classificação de finalidade o tamanho do ranking foi definido como cinco (número máximo de classes do problema), e a pontuação mínima para ranqueamento foi definida em um. No caso do problema de classificação de categoria o tamanho do ranking foi fixado em dez e a pontuação mínima para ranqueamento em três.

Além dos parâmetros oferecidos pelo classificador, foram estabelecidos dois parâmetros de integração, um para cada problema de classificação abordado nos experimentos.

Na classificação de finalidade, o parâmetro de integração indica como a aplicação integradora deverá tratar a saída do classificador, caso não seja estabelecida uma classe atribuída ao documento. Este parâmetro indica se a finalidade Ensino deverá ser atribuída como padrão na ausência de classificação.

Para o problema de classificação de categoria, o parâmetro de integração está relacionado com a entrada do classificador. Indicando, se o conjunto de classes elegíveis para classificação da amostra deverá ser enviado integralmente com todas as categorias possíveis, ou se o conjunto deverá ser enviado somente com as classes da respectiva árvore referente ao tipo de registro contábil analisado (crédito ou débito).

Estabelecidos os parâmetros fixos para cada problema de classificação abordado (tamanho do ranking e pontuação mínima para ranqueamento), cinco parâmetros binários foram experimentados nas suas 32 possíveis combinações. para se estabelecer a melhor configuração para cada problema de classificação. A Seção 6.5 apresenta detalhadamente estes parâmetros e resultados obtidos.

6.5 Resultados do Método de Classificação

As Tabelas 12 e 13 apresentam o resumo dos experimentos (EXP) realizados com o classificador proposto, destacando resultados com a utilização de todos os parâmetros binários de configuração, o melhor resultado considerado, os melhores resultados sem utilizar parâmetro de integração, o pior resultado registrado e, por fim, o resultado sem a utilização dos os parâmetros binários de configuração.

A última coluna das tabelas é dedicada ao indicador de erros (ERR), que demonstra a quantidade de documentos sem classe atribuída pelo classificador. Nestes casos, não é possível contabilizar o resultado na matriz de confusão da classe. Portanto, a indicação de erros anula resultados de acurácia superior e é utilizada como desempate na escolha do pior resultado.

A colunas de acurácia média (ACC) demonstram o desempenho do classificador nos experimentos, destacando valores invalidados pela ocorrência de erro. As demais colunas registram com *checkmark* (✓) a ativação de seus respectivos parâmetros.

A coluna *"PI-E"* na Tabela 12 significa *Parâmetro de Integração - Ensino como padrão.* Enquanto na Tabela 13 a coluna *"PI-S"* significa *Parâmetro de Integração - Separar classes*

por tipo. As demais colunas possuem os mesmos significados em ambas tabelas, representando parâmetros de personalização disponibilizados pelo classificador, detalhados na Seção 5.4: forçar a classificação (FRC), utilizar critérios de desempate (UCD), normalizar strings (NRM) e utilizar somente letras minúsculas (SLM).

Tabela 12 – Resultados para a Classificação de Finalidade

EXP	FRC	UCD	PI-E	NRM	SLM	ACC	ERR
Todos parâmetros	✓	✓	✓	✓	✓	87.10%	0
Melhor resultado	–	✓	✓	–	✓	**88.45**%	0
Sem PI-E_1	–	✓	–	–	✓	~~88.44%~~	1
Sem PI-E_2	✓	✓	–	–	✓	88.43%	0
Pior resultado	–	–	✓	–	–	83.73%	0
Nenhum parâmetro	–	–	–	–	–	~~90.08%~~	732

Fonte: Produzida pelo autor.

O parâmetro de forçar a classificação (FRC) evita que documentos não sejam classificados, mas quando ativo pode interferir na acurácia do classificador, pois os registros com erro são ignorados na matriz de confusão, o que pode distorcer o valor atribuído aos acertos do algoritmo. Diante do exposto, os melhores resultados foram selecionados preferencialmente na ausência deste parâmetro. Em contrapartida, a parametrização para utilizar critérios de desempate (UCD), demonstrou ser eficiente na minimização da quantidade de documentos sem classe atribuída, sendo na maioria das vezes suficiente para zerar o indicador de erros.

Tabela 13 – Resultados para a Classificação de Categoria

EXP	FRC	UCD	PI-S	NRM	SLM	ACC	ERR
Todos parâmetros	✓	✓	✓	✓	✓	99.19%	0
Melhor resultado	–	✓	✓	✓	✓	**99.19**%	0
Sem PI-S	–	✓	–	✓	✓	98.92%	0
Pior resultado	–	✓	–	–	–	98.44%	4
Nenhum parâmetro	–	–	–	–	–	98.54%	526

Fonte: Produzida pelo autor.

O parâmetro de integração (PI-S) do problema de classificação de categoria demonstrou-se eficiente, pois a utilização de classes separadas por tipo aumentou a acurácia do classificador. Outro benefício observado com a ativação deste parâmetro, foi a diminuição do tempo de execução pela metade.

A utilização de textos padronizados em letras minúsculas (SLM) melhorou o desempenho do classificador em todos os testes, nota-se que os melhores resultados de acurácia nas Tabelas 12 e 13 apresentam este parâmetro como ativo. Entretanto, a utilização de textos normalizados (NRM), apresentou comportamento distinto de acordo com cada problema de classificação. No problema das finalidades, a utilização de textos não normalizados contribuiu para melhora do desempenho do classificador em mais de um ponto percentual. Enquanto no problema das categorias, os melhores resultados foram obtidos com a utilização de textos normalizados.

6.6 Comparação de Desempenho na Classificação de Finalidade

O desempenho do classificador proposto foi de 88.45% de acurácia na classificação de finalidade. Além do bom desempenho registrado para este problema de classificação, nota-se que a diferença entre o melhor e o pior desempenho do classificador, foi de apenas 4,72 pontos percentuais. Nota-se também que o parâmetro de integração (PI-E) é irrelevante quanto ao desempenho do classificador, pois sua ausência diminuiu a acurácia em apenas dois centésimos.

Observando-se o desempenho dos algoritmos base de referência descritos na Tabela 14, nota-se que somente o SVM conseguiu superar o limiar dos 88%, registrando 88,73% de acurácia. Seguido de dois dos três métodos com características de aprendizagem preguiçosa, k-NN com 87,72% de acurácia e K^* com 87,66%.

Tabela 14 – Resultados dos Algoritmos Base de Referência na Classificação de Finalidade

Algoritmo	Implementação utilizada no Weka	Acurácia
k-NN	lazy.IBk	87,72%
K^*	lazy.Kstar	87,66%
LWL	lazy.LWL	72,86%
NB	bayes.NaiveBayes	68,48%
DT	trees.J48	81,78%
SVM	LibSVM – utilizando *Kernel Linear*	**88,73**%

Fonte: Produzida pelo autor.

Estes resultados posicionam o classificador proposto, como o segundo melhor desempenho geral para este problema de classificação.

6.7 Comparação de Desempenho na Classificação de Categoria

O desempenho do método proposto foi de 99,19% de acurácia neste problema de classificação, superando o SVM em 10,46 pontos percentuais, estabelecendo o classificador proposto como melhor método para a classificação de categoria.

A Tabela 15 demostra os o resultados obtidos pelos algoritmos base de referência. Considerando apenas estes algoritmos, o SVM apresentou melhor resultado, obtendo 86,70% de acurácia, também sendo seguido por dois dos três algoritmos preguiçosos, o K^* com 86,33% de acurácia e k-NN com 85,81%.

6.8 Análise e Discussão dos Resultados

A análise de desempenho do classificador, demonstra eficácia do método proposto ao registrar níveis de acurácia superiores a 88%. Enquanto a eficiência do classificador foi comprovada com o registro de melhor desempenho ou segundo melhor desempenho obtido em um problema de classificação, sendo que o classificador proposto superou o SVM em mais de dez pontos percentuais quando obteve o melhor desempenho, e ficou menos de três décimos atrás do SVM quando obteve o segundo melhor desempenho.

Tabela 15 – Resultados dos Algoritmos Base de Referência na Classificação de Categoria

Algoritmo	Implementação utilizada no Weka	Acurácia
k-NN	lazy.IBk	85,81%
K^*	lazy.Kstar	86,33%
LWL	lazy.LWL	35,14%
NB	bayes.NaiveBayes	77,32%
DT	trees.J48	82,94%
SVM	LibSVM – utilizando *Kernel Linear*	**86,70%**

Fonte: Produzida pelo autor.

A simplicidade do método proposto é evidenciada pela facilidade de implementação do classificador, e pela facilidade de realização da tarefa de classificação de texto, devido à ausência da necessidade de pré-processamento ou construção de modelos de aprendizagem. Ressalta-se também a estabilidade do classificador, ao registrar variação inferior a cinco pontos percentuais, analisando-se os desempenhos obtidos no melhor e no pior caso de um mesmo problema de classificação. Da mesma forma, fica evidente a genericidade do classificador, devido sua capacidade de utilização em diferentes tipos problemas de classificação de texto.

Além da simplicidade e genericidade destacada, o classificador proposto apresenta versatilidade, pois a capacidade de parametrização do seu funcionamento, em conjunto com a produção de saída interpretável, permitem que este método de classificação seja facilmente integrado à sistemas legados. Por fim, conclui-se que o método de classificação desenvolvido é uma solução *soft computing* adequada para realização da classificação automática dos registros financeiros e contábeis.

6.9 Considerações Finais sobre a Avaliação do Método de Classificação

Este capítulo apresentou a avaliação do método de classificação estabelecido e posicionou-o em relação aos seus pares.

A solução de classificação de texto estabelecida no Capítulo 5 e validada no Capítulo 6, mostrou-se eficiente para na atribuição de categorias discriminativas e atribuição de finalidades qualitativas. Possibilitando a caracterização dos registros financeiros e contábeis gerados pela execução orçamentária da universidade. Propõe-se então, que esta solução seja a ferramenta utilizada no pré-processamento dos dados para a construção das ferramentas de visualização da informação.

7 Solução Proposta

A partir dos conceitos e métodos estabelecidos nos capítulos anteriores, este capítulo dedica-se à apresentação da proposta de solução para o problema tema deste trabalho. Retomando o cenário de testes, apresentando as ferramentas de visualização implementadas, provendo informações sobre a implantação da solução e, por fim, apresentando discussão sobre avaliação da solução.

7.1 Cenário de Implantação

Conforme descrito na Seção 2.2, a Faculdade de Engenharia – Campus de Bauru (FEB), foi a unidade escolhida para implantação experimental do protótipo desenvolvido neste trabalho. A execução orçamentária da Faculdade de Engenharia é realizada de forma distribuída em centros de custo hierarquicamente organizados, sendo que um centro de custo não necessariamente representa um componente do organograma da instituição. Considerando a ótica da gestão administrativa, o conceito de centro de custo é uma abstração do organograma da faculdade.

Figura 6 – Proposta de organização dos centros de custo na universidade.

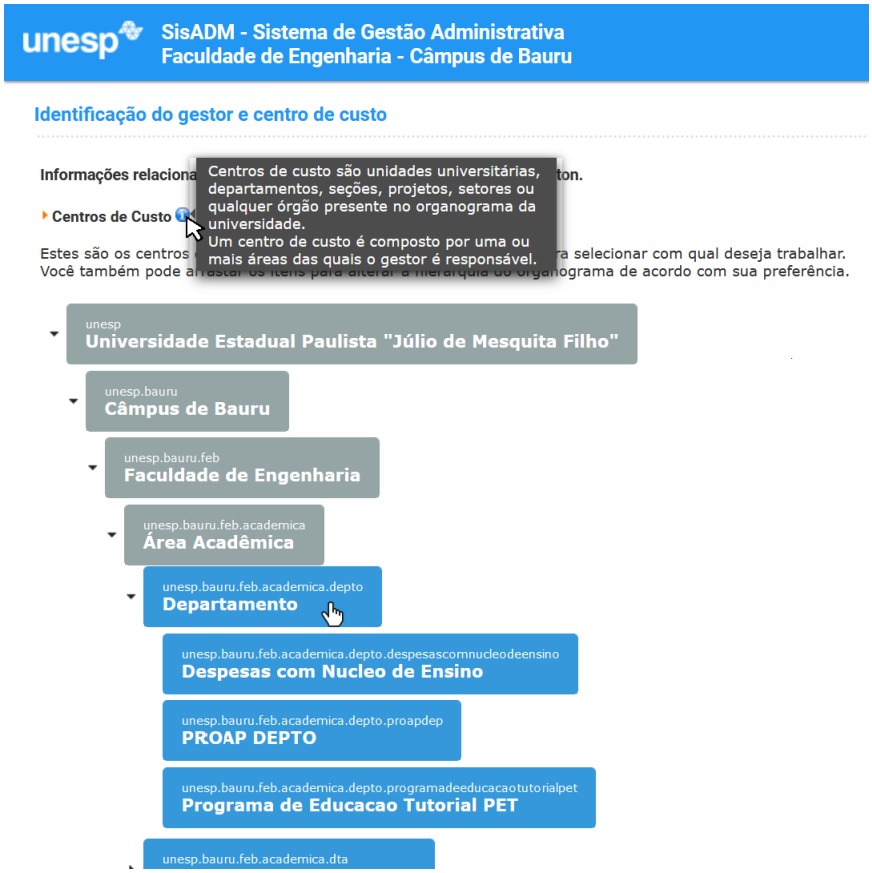
Fonte: Produzida pelo autor.

A Figura 6 representa a organização hierárquica dos centros de custo proposta neste trabalho e implantada no sistema SisADM. Nota-se na figura o centro de custo chamado *"Departamento"*, o qual foi utilizado como exemplo nas imagens. Destacando-se também o centro de custo *"Folha de Pagamento FEB"*, no qual estão vinculados os registros diretamente vinculados à folha de pagamento, ficando este isolado e não sendo explorado neste estudo.

Os dados utilizados nas ferramentas de visualização apresentadas neste capítulo, consistem nos 3452 registros do ano de 2019 do *Data Mart* estabelecido para este trabalho. Destes, constam 244 (7,07%) registros diretamente relacionados à folha de pagamento ou despesas com pessoal. Sendo os demais 3208 (92,93%) registros estão relacionados com as outras atividades inerentes da execução orçamentária na FEB.

A Figura 7 apresenta a visualização disponível para os usuários envolvidos no projeto, com destaque para o balão informativo descrevendo o conceito de centro de custo aplicado no sistema, destacando-se também a expansão do centro de custo *"Departamento"*, com a exibição de seus centros de custo subordinados.

Figura 7 – Informação sobre centros de custo e expansão do centro de custo *"Departamento"*.



Fonte: Produzida pelo autor.

Os retângulos que representam os centros de custo nas Figuras 6 e 7, possuem diferenciação de cores entre cinza e azul. A coloração cinza é utilizada para representar centros de custo intermediários aos quais o usuário não está relacionado como gestor. Enquanto a coloração azul, indica que o usuário tem privilégios de gestor neste centro de custo.

7.2 Aplicação do Processo de Classificação de Texto

Esta seção aborda a execução do processo de classificação de texto nos registros do *Data Mart*, como forma de pré-processamento de dados para as ferramentas de visualização da informação.

Conforme processo estabelecido na Seção 3.6, destacam-se aqui algumas das ações e escolhas realizadas nas três fases do processo de classificação de texto (pré-processamento, treinamento, e validação). Contudo, antes de abordar a execução do processo, é necessário escolher um método de classificação.

7.2.1 Escolha do Método de Classificação

Fundamentado por Santos e Merschmann (2020) destacando que, apesar da mineração de dados e o aprendizado de máquina possuem várias técnicas que podem apoiar a análise de texto, encontrar a técnica mais adequada para cada problema é uma tarefa dispendiosa, pois requer conhecimento técnico avançado, exaustivos experimentos computacionais e, consequentemente, tempo.

Fundamentado também pela constatação Liu et al. (2017), sobre métodos os *Deep Learning* precisarem de muito mais dados de treinamento para se obter um desempenho vantajoso em relação a outros métodos. De modo que seria necessário o conjunto de dados experimentais ter grande número de instâncias de treinamento para cada classe elencada.

Considerando os problemas de classificação de finalidade e categoria, possuem classes altamente desbalanceadas no conjunto de dados experimental. Sendo que Xiang e Zheng (2018) comprovam a piora da precisão na classificação quando o conjunto de dados não é balanceado. Considerando ainda que para o problema de classificação de categorias, não é viável a aplicação da técnica de balanceamento amostral SMOTE (do inglês, *Synthetic Minority Oversampling Technique*) proposta por Chawla et al. (2002), pois existem muitas classes sem registros de exemplo ou com apenas um registro amostral.

Considerando, por fim, que a solução de classificação de texto estabelecida no Capítulo 5 e validada no Capítulo 6, demonstrou ser eficiente na tarefa de classificação automática de texto. Optou-se então pela não utilização de métodos de ML ou DL para classificação dos registros financeiros e contábeis.

Estabelece-se então, que a solução de classificação apresentada neste trabalho, será a ferramenta utilizada no pré-processamento dos dados para a construção das ferramentas de visualização da informação.

7.2.2 Processo de Classificação de Texto

Para cada registro financeiro e contábil presente no *Data Mart*, origina-se um documento textual com título e conteúdo descritivo. O título do documento é gerado pelo sistema SisADM, frequentemente referindo-se à origem do registro ou ao seu ato administrativo vinculado. O conteúdo descritivo do registro é formado pelo campo de descrição, adicionado das informações

presentes nos demais campos opcionais do registro. A fase de pré-processamento tem como principal objetivo realizar a modelagem deste documento e realizar a seleção de características que serão utilizadas para classificação.

O modelo de representação vetorial *Bag of Words* (BoW) foi escolhido para construção da matriz de termos do documento. A partir deste vetor, tokens *unigram (1-gram)* são disponibilizados para a comparação de similaridade, que executa de forma implícita tratamento equivalente aos procedimentos *stemming* e *lemmatisation*.

São realizados os procedimentos *case folding*, para padronização de texto em caixa baixa, e *cleaning*, para remoção de caracteres especiais, dígitos, sinais de pontuação e acentuação. Realiza-se também o procedimento *length filtering*, para remoção de palavras pequenas. Sendo este último realizado de forma dinâmica pela solução de classificação desenvolvida, na qual consideram-se válidas as palavras cujo tamanho supere ao menos: i) um valor mínimo definido por parâmetro; ou, ii) o menor tamanho de palavra-chave envolvida na classificação.

A remoção de palavras irrelevantes como artigos e preposições (*stopwords removal*) não é realizada, pois estes termos podem estar presentes em palavras-chave compostas.

Conforme descrito no Capítulo 5, o classificador desenvolvido não necessita de treinamento, pois todo aprendizado está incutido nos dicionários de palavras-chave criados para cada classe apresentada ao classificador.

Por fim, a validação deste método de classificação foi realizada utilizando-se a métrica acurácia média.

7.3 Ferramentas de Visualização da Informação

Inspirado no modelo de visualização estático criado manualmente (visto na Figura 3, que foi apresentada no Item 2.4.1) e considerando a questão de como a Informática poderia contribuir para melhoria da compreensão humana sobre os dados financeiros e contábeis, foram criadas ferramentas de visualização da informação que possuem comportamento interativo, permitindo aos usuários realizar a exploração da execução orçamentária.

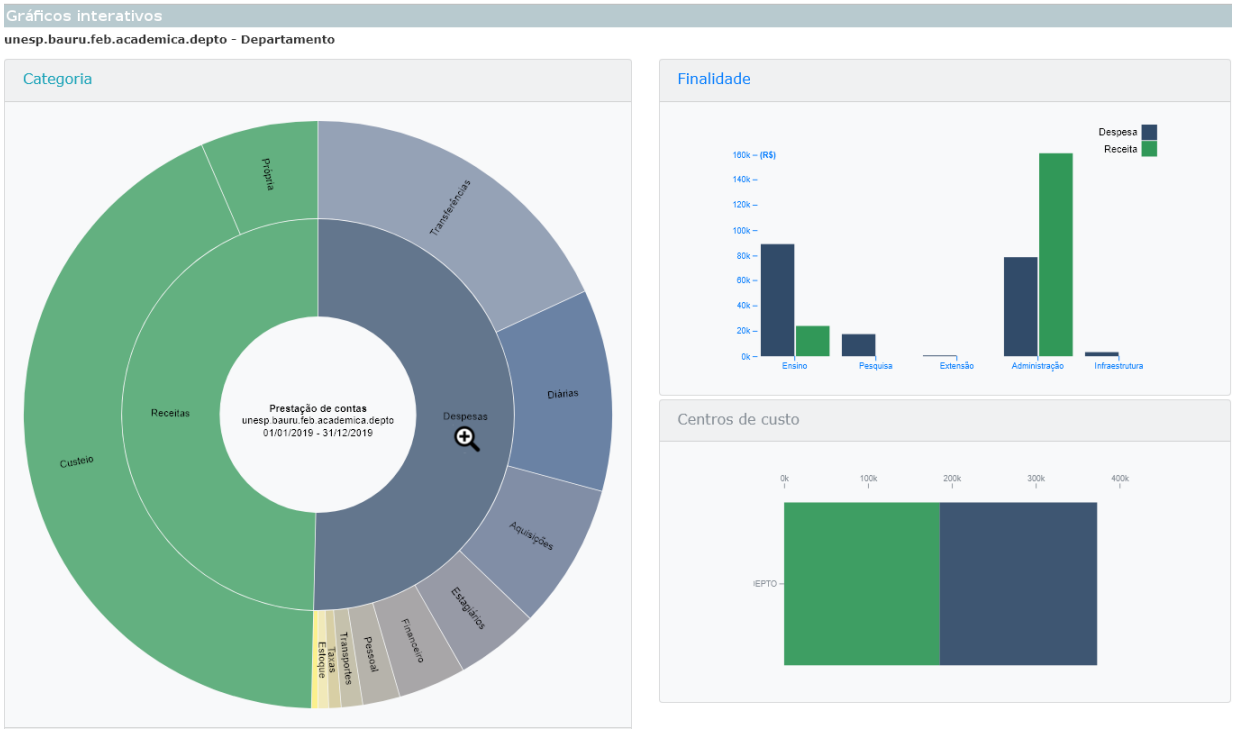
A implementação das ferramentas de visualização foi realizada seguindo padrões *Web*, com a utilização de linguagem de marcação *Hypertext Markup Language* (HTML), linguagem de folha de estilo *Cascading Style Sheets* (CSS) e a biblioteca *D3.js – Data-Driven Documents* (D3). Enquanto HTML e CSS são conceitos amplamente consolidados, a biblioteca D3 não é tão conhecida. Segundo Bostock (2021), D3.js é uma biblioteca *JavaScript* para manipulação de documentos baseados em dados, uma biblioteca independente e moderna, a qual irá lhe ajudar a dar vida aos dados.

7.3.1 *Dashboard* Interativo

A Figura 8 apresenta o *dashboard* interativo elaborado neste trabalho. Esta ferramenta de visualização combina três ferramentas gráficas que funcionam de forma integrada. O gráfico de rosca apresenta informações relacionadas às categorias dos registros financeiros e contábeis,

este gráfico permite a navegação em profundidade conforme a estrutura hierárquica presente nas categorias relacionadas. O cursor denominado *zoom-in*, que possui formato de lupa com o símbolo "+", representa esta possibilidade de navegação em profundidade na categoria indicada.

Figura 8 – Visão geral do *dashboard* interativo.

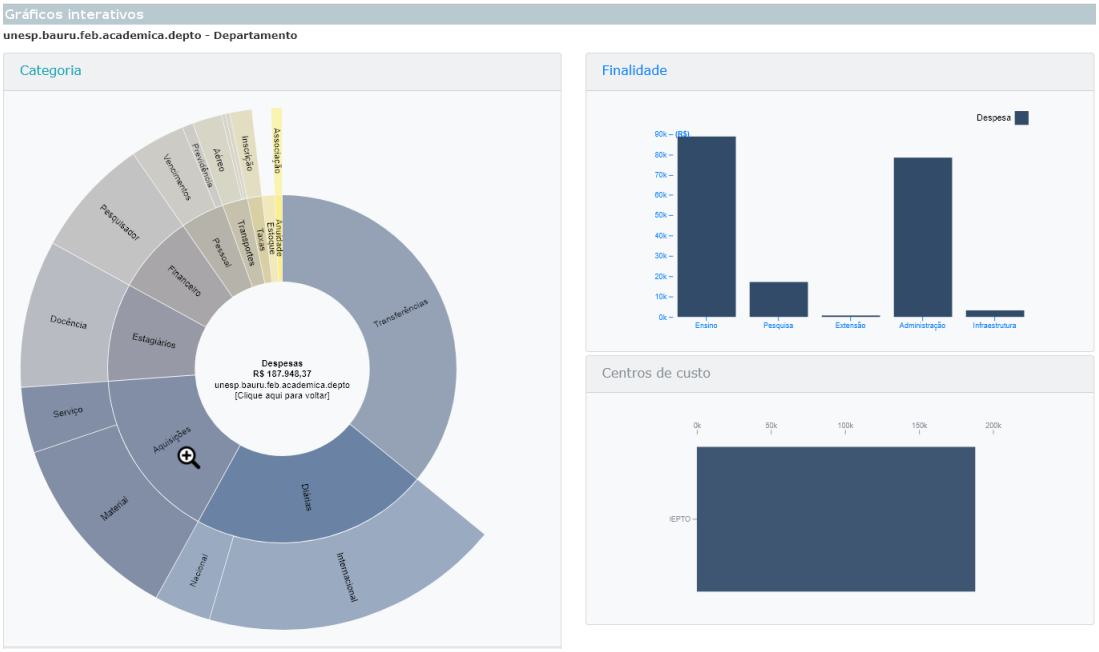


Fonte: Produzida pelo autor.

Conforme realiza-se a navegação em profundidade, os demais gráficos do *dashboard* são atualizados para corresponder com os dados da categoria selecionada no gráfico de rosca. O gráfico de barras verticais agrupadas, apresenta informações referentes aos valores totais de registros de débito e/ou crédito, informações agrupadas de acordo com as finalidades atribuídas aos lançamentos. Já o gráfico de barras horizontais empilhadas, representa a distribuição de lançamentos de débitos e/ou créditos para cada centro de custo que compõe a estrutura organizacional observada. As informações apresentadas na Figura 8 correspondem a visão inicial, tanto de débitos quanto de créditos, somente para o centro de custo *"Departamento"* e ignorando dados de seus centros de custo subordinados.

As Figuras 9 e 10 complementam a representação da navegação em profundidade. Partindo da visão exibida na Figura 8, entra-se na categoria de *"Despesas"* na Figura 9, depois entra-se na categoria de *"Aquisições e contratações"* na Figura 10. Nota-se na sequência de figuras a alteração dos gráficos de finalidade e centro de custo relacionados com os registros da categoria selecionada.

Figura 9 – *Dashboard* após navegação em profundidade para o item *"Despesas"*.

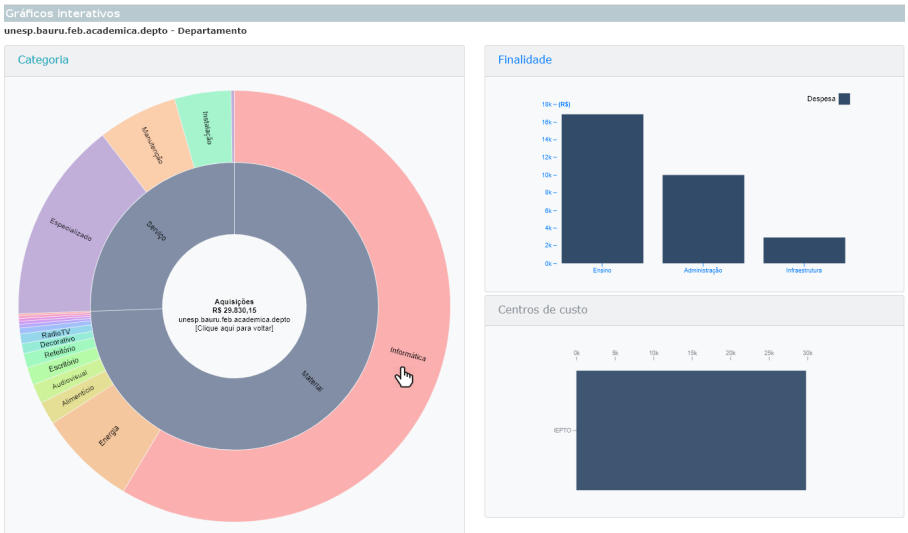


Fonte: Produzida pelo autor.

No gráfico de rosca apresentado na Figura 10, o cursor destacado em formato de mão representa que a categoria observada não possui subcategorias. Mas ainda seria possível a navegação para níveis mais detalhados como *"Aquisição de materiais"* e *"Contratação de serviços"*, sendo estes os níveis mais profundos de navegação do gráfico.

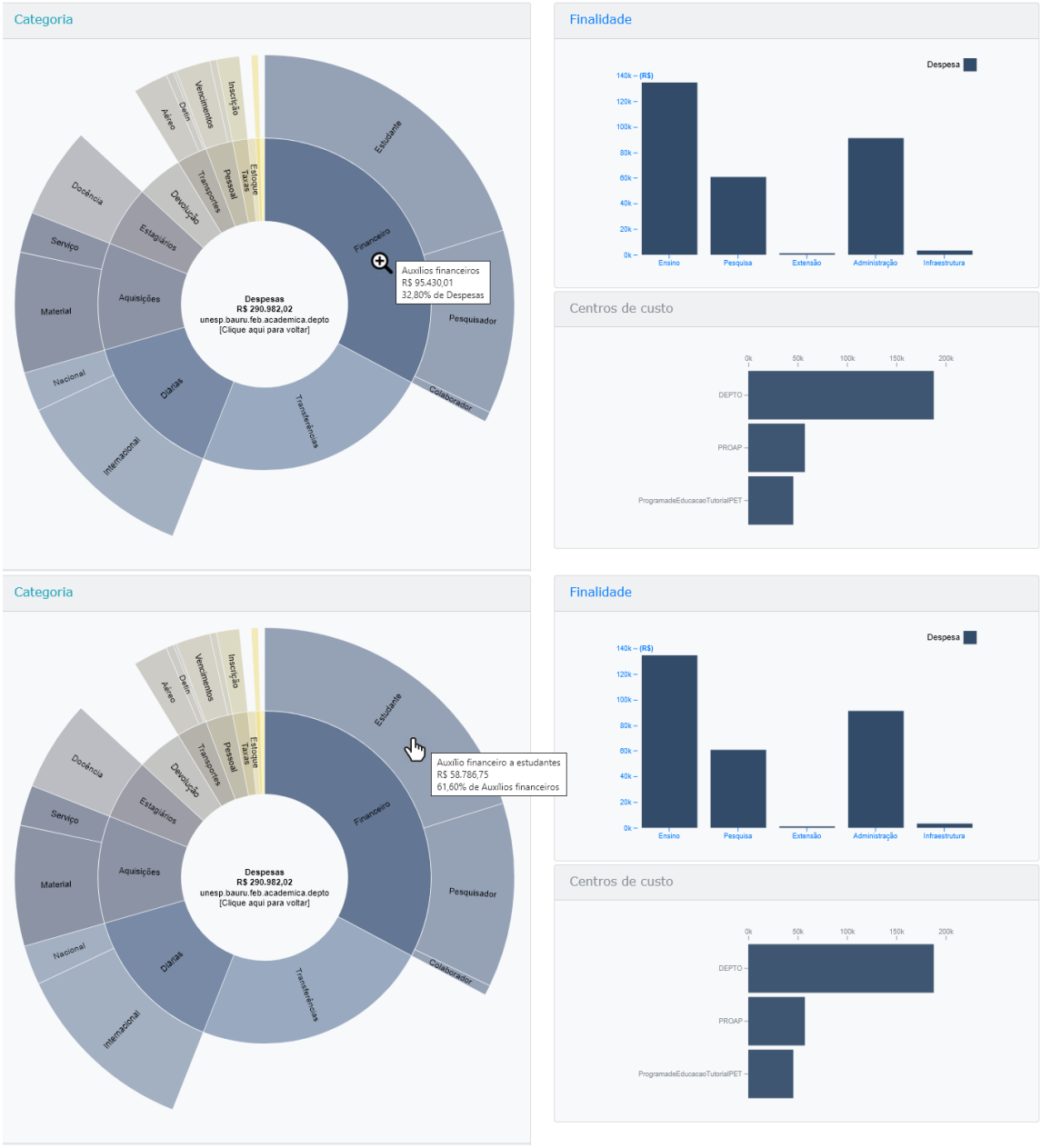
Os gráficos exibidos na Figura 11 apresentam o detalhamento de informações, ao se posicionar o cursor sobre a categoria de interesse. Neste gráficos constam informações de despesas do centro de curso *"Departamento"* e seus dois centros de custo subordinados, *"PROAP"* e *"ProgramaEducacaoTutorialPET"*.

Figura 10 – *Dashboard* após navegação em profundidade para o item *"Aquisições"*.



Fonte: Produzida pelo autor.

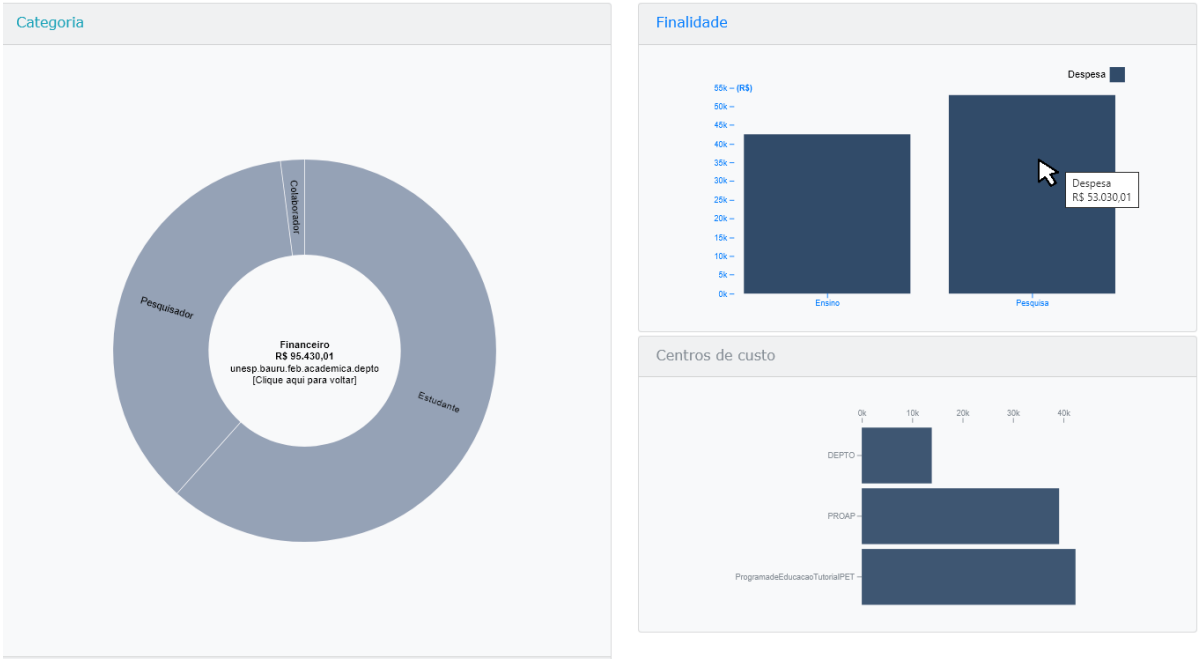
Figura 11 – Informações detalhadas pelo posicionamento do cursor em *"Categoria"*.



Fonte: Produzida pelo autor.

A Figura 12 complementa a representação do detalhamento de informações, ao se posicionar o cursor em algum item de interesse.

Figura 12 – Informações detalhadas pelo posicionamento do cursor em *"Finalidade"*.



Fonte: Produzida pelo autor.

Além da parte visual com as ferramentas gráficas, o *dashboard* interativo possui um painel de controle, que permite realizar a filtragem dos dados de interesse à serem apresentados graficamente. A Figura 13 apresenta o painel de controle do *dashborad*, com destaque para o ano de 2019 como período selecionado. Destacam-se também as indicações com cursor *pointer* (em formato de mão apontando), que ressaltam demais funcionalidades para filtragem de dados da exibição. O campo *"Considerar TODOS os registros"* permite ao usuário selecionar quais são as categorias de interesse para serem exibidas na visualização, a indicação selecionada na imagem consiste na exibição de todas as categorias. As outras duas funcionalidades apontadas, indicam a possibilidade de consideração (ou não) dos centros de custo subordinados e a possibilidade da exibição dos dados em tabela.

Figura 13 – Painel de controle do *dashboard* interativo.

Fonte: Produzida pelo autor.

Para melhor aproveitamento de espaço e, consequentemente, melhor representação da tabela de registros que originam as representações visuais do *dashboard*, a página seguinte dedica-se à exibição da Figura 14 utilizando a orientação paisagem. Esta tabela pode ser acessada ao se clicar no link *"Visualizar tabela de registros dos gráficos..."* disponível no painel de controle da ferramenta.

Figura 14 – Tabela de registros utilizados na visualização *Dashboard*.

Relação de registros de prestação de contas - Departamento (+ centros de custo subordinados)

Período selecionado: Janeiro/2019 até Dezembro/2019. [48 registros]

Categoria raíz selecionada: despesas.financeiro.

Data	Tipo	Título	Valor (R$)	Finalidade	Categoria	Centro de custo	
01/03/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO A PESQUISADOR 2019.	R$ 1.785,94	Pesquisa	Auxílio financeiro a pesquisadores	Departamento	AUXÍLIO FINANCEIRO A PESQUISA SOLICITAÇÃO DE EMPENHO ANEXA (DECRETO ESTADUAL Nº 64.078, DE 2 AUTORIZADAS PELO ORDENADOR DA
01/03/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO A PESQUISADOR 2019.	R$ 3.395,33	Pesquisa	Auxílio financeiro a pesquisadores	Departamento	AUXÍLIO FINANCEIRO A PESQUISAD EMPENHO ANEXA (OFÍCIO Nº 12/201 Nº 64.078, DE 21 DE JANEIRO DE 20 ORDENADOR DA DESPESA NO PROCES
13/03/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO A PESQUISADOR 2019.	R$ 2.261,37	Pesquisa	Auxílio financeiro a pesquisadores	Departamento	AUXÍLIO FINANCEIRO A PESQUISADO CONFORME SOLICITAÇÃO DE EMPE ESTADUAL Nº 64.078, DE 21 DE J AUTORIZADAS PELO ORDENADOR DA
02/05/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO A PESQUISADOR 2019.	R$ 1.856,32	Pesquisa	Auxílio financeiro a pesquisadores	Departamento	AUXÍLIO FINANCEIRO A PESQUISADO SOLICITAÇÃO DE EMPENHO ANEXA (O THE INTERNATIONAL WORKSHOP GUIMARÃES/PORTUGAL). FORAM ATE DE JANEIRO DE 2019 - DOE DE 22 D DESPESA NO PROCESSO EM USO.
07/06/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO A PESQUISADOR 2019.	R$ 799,99	Pesquisa	Auxílio financeiro a pesquisadores	Departamento	AUXÍLIO FINANCEIRO A PESQUISADO DE EMPENHO ANEXA (OFÍCIO Nº 37, ESTADUAL Nº 64.078, DE 21 DE J AUTORIZADAS PELO ORDENADOR DA
19/08/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO A PESQUISADOR 2019.	R$ 3.783,63	Pesquisa	Auxílio financeiro a pesquisadores	Departamento	AUXÍLIO FINANCEIRO A PESQUISA SOLICITAÇÃO DE EMPENHO ANEXA (DECRETO ESTADUAL Nº 64.078, DE 2 AUTORIZADAS PELO ORDENADOR DA
28/01/2019	Despesa	PROAP DEPTO	R$ 3.500,00	Pesquisa	Auxílio financeiro a pesquisadores	PROAP DEPTO	CONVÊNIO CAPES PROAP Nº817737/ PESQUISADOR - PROF FORAM ATENDIDAS ÀS DISPOSIÇÕES DOE DE 22 DE JANEIRO DE 2019. DES EM USO.
05/02/2019	Despesa	PROAP DEPTO	R$ 3.422,40	Pesquisa	Auxílio financeiro a pesquisadores	PROAP DEPTO	CONVÊNIO CAPES PROAP Nº 817737 PESQUISADOR - PROF ÀS DISPOSIÇÕES DO DECRETO ESTAD DE 2019. DESPESAS AUTORIZADAS PE
21/02/2019	Despesa	PROAP DEPTO	R$ 1.577,60	Pesquisa	Ajuda de custo a colaboradores eventuais	PROAP DEPTO	CONVÊNIO CAPES PROAP nº 817737/2 DESPESA COM LIBERAÇÃO DE AJ PARTICIPAÇÃO EM ATIVIDADES DESE DOCUMENTAÇÃO JUNTADA AO PROC ESTADUAL Nº 64.078, DE 21 DE J AUTORIZADAS PELO ORDENADOR DA
26/04/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO ALUNO 2019.	R$ 463,40	Pesquisa	Auxílio financeiro a estudantes	PROAP DEPTO	PROAP DEP TESOURO. AUXÍLIOS - RELACIONADO COMO CREDOR/BENE LIBERAÇÃO: MEDIANTE CHEQUE N DISPOSIÇÕES DO DECRETO ESTADUA 2019. DESPESAS AUTORIZADAS PELO
02/05/2019	Despesa	DESPESAS COM AUXILIO FINANCEIRO ALUNO 2019.	R$ 463,40	Pesquisa	Auxílio financeiro a estudantes	PROAP DEPTO	AUXÍLIO FINANCEIRO A ESTUDAN LIBERAÇÃO(ES) DE BOLSA(S) CREDOR/BENEFICIÁRIO. FORMA DE L ATENDIDAS ÀS DISPOSIÇÕES DO DEC DE JANEIRO DE 2019. DESPESAS AUT
05/06/2019	Despesa	PROAP DEPTO	R$ 926,80	Pesquisa	Auxílio financeiro a estudantes	PROAP DEPTO	CONVÊNIO CAPES PROAP nº 817737/ LIBERAÇÃO(ES) DE AUXÍLIO(S) AO(S DE LIBERAÇÃO: MEDIANTE CRÉDITO DISPOSIÇÕES DO DECRETO ESTADUA 2019. DESPESAS AUTORIZADAS PELO
							CONVÊNIO CAPES PROAP nº 817737 PESQUISADOR - PROF

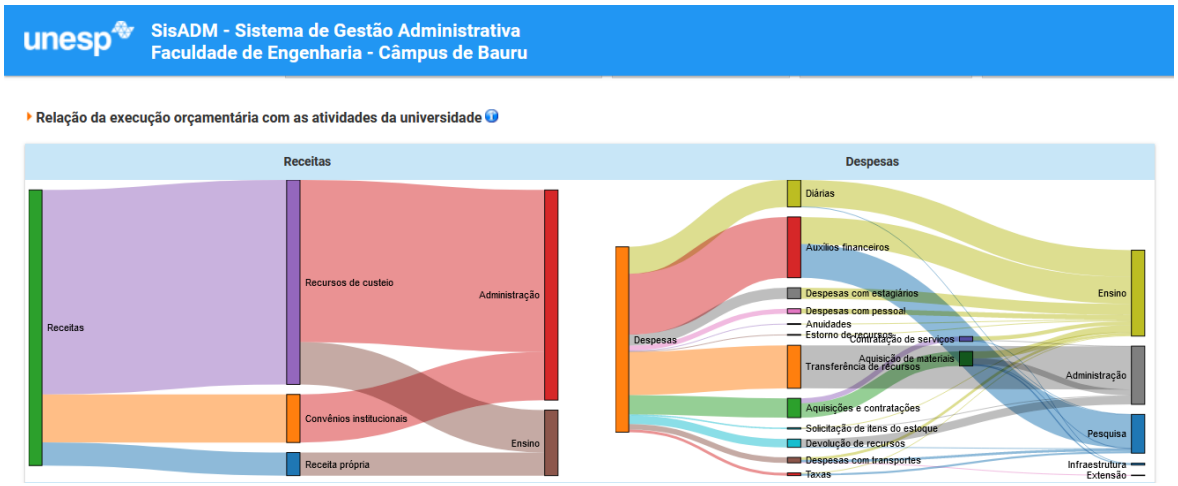
Fonte: Produzida pelo autor.

7.3.2 Diagrama de Sankey

O Diagrama de Sankey é utilizado neste trabalho como ferramenta de visualização para demonstrar o relacionamento entre os domínios Categoria e Finalidade. No intuito de representar a execução orçamentária do centro de custo, em relação as atividades da universidade.

A Figura 15 apresenta a tela inicial da funcionalidade de acompanhamento da execução orçamentária do ano atual. O gráfico de *"Receitas"* relaciona toda a entrada de recursos no centro de custo, com as finalidades as quais estes recursos foram destinados. Enquanto o gráfico de *"Despesas"* relaciona os gastos do centro de custo, com as finalidades institucionais na qual foram empregados. Os retângulos verticais representam as categorias e finalidades relacionadas com o emprego dos recursos financeiros. Enquanto os caminhos que interligam os retângulos, expressam a relação entre categorias e finalidades.

Figura 15 – Página de acompanhamento da execução orçamentária do Exercício Vigente.

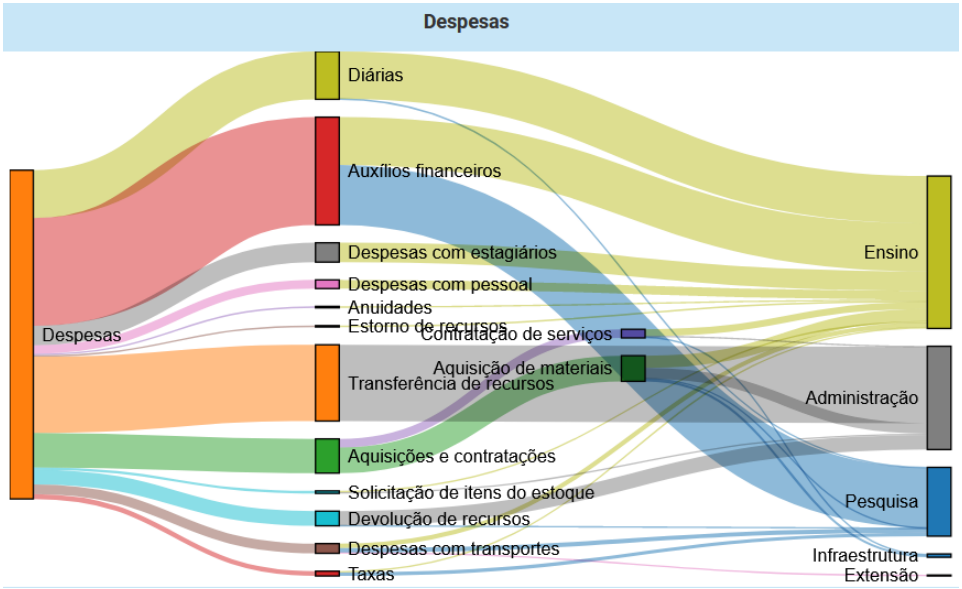


Fonte: Produzida pelo autor.

Além de exprimir a relação entre categorias e finalidades institucionais, os diagramas construídos permitem manipulação e interatividade. Os retângulos presentes nos diagramas podem ser movidos livremente, este comportamento é um tratamento de oclusão que permite o ajuste do digrama de acordo com a melhor visualização para o usuário.

A Figura 16 apresenta a visualização original do gráfico de *"Despesas"* inicialmente visto na Figura 15.

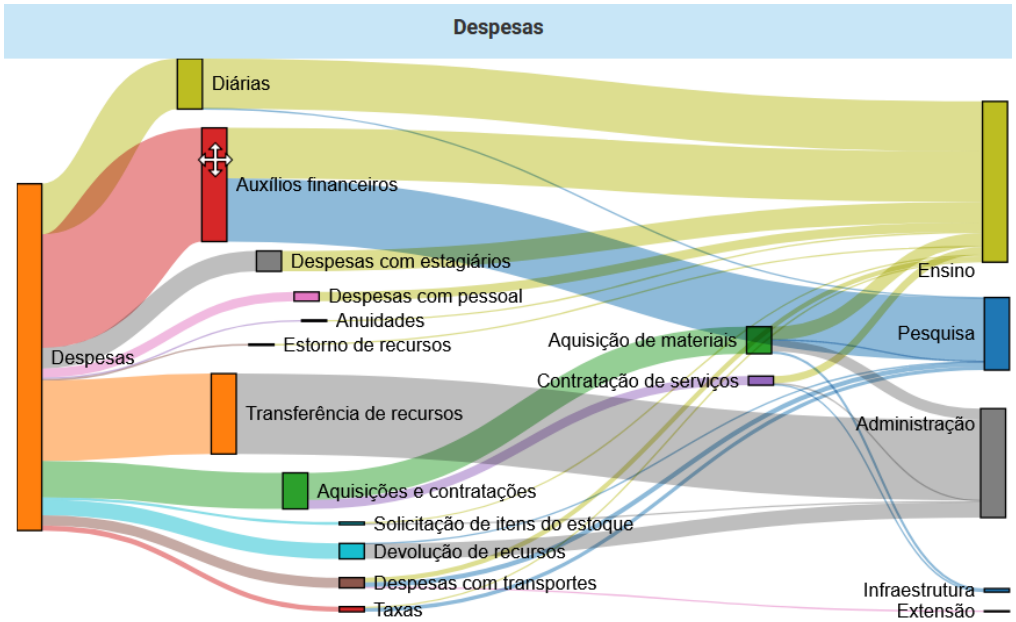
Figura 16 – Detalhes no gráfico de *"Despesas"*.



Fonte: Produzida pelo autor.

Ao passo que a Figura 17 apresenta a reorganização do gráfico da Figura 16, com destaque para o cursor de movimentação exibido sobre o retângulo referente a categoria *"Auxílios financeiros"*.

Figura 17 – Gráfico de *"Despesas"* reorganizado.

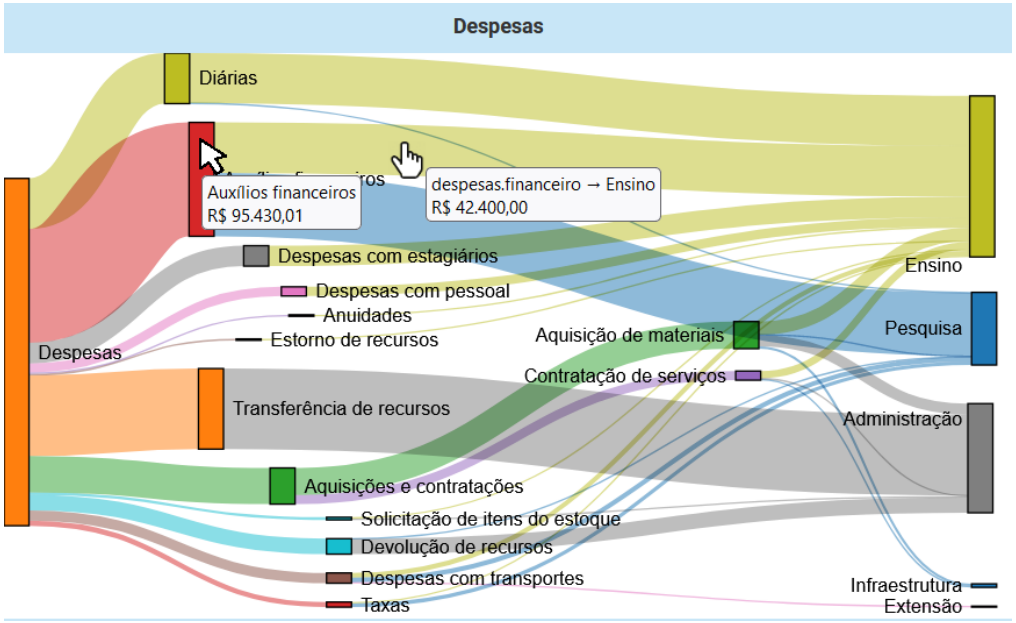


Fonte: Produzida pelo autor.

De forma semelhante a ferramenta *Dashboard* Interativo, os gráficos exibidos na ferramenta Diagrama de Sankey também apresentam o detalhamento de informações, ao se posicionar o cursor sobre o item de interesse. A Figura 18 exibe o detalhamento geral da categoria *"Auxílios financeiros"*, mostrando também informação específica da relação desta categoria com a finalidade

"Ensino"

Figura 18 – Detalhamento de informações sobre *"Auxílios financeiros"*.



Fonte: Produzida pelo autor.

Outro recurso de interatividade disponível neste diagrama, é a exibição dos registros que estabelecem o valor do *link*. A Figura 19 apresenta a tabela disponibilizada para o usuário ao se clicar no *link* entre *"Auxílios financeiros"* e *"Ensino"*.

Figura 19 – Tabela de registros utilizados na visualização *Diagrama de Sankey*.

Relação de registros de prestação de contas - Departamento (+ centros de custo subordinados)

Período selecionado: Janeiro/2019 até Dezembro/2019. [10 registros]

Categoria raíz selecionada: despesas.financeiro.

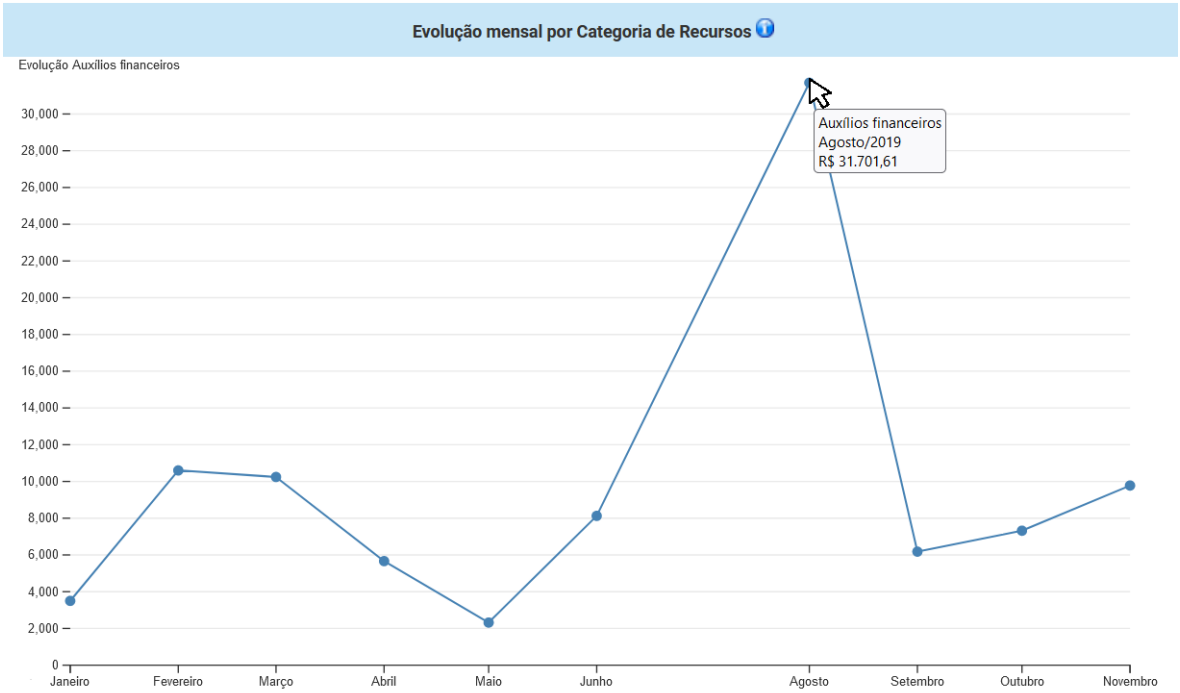
Data	Tipo	Título	Valor (R$)	Finalidade	Categoria	Centro de custo
26/02/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 5.600,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
27/03/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 2.800,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
05/04/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 5.200,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
24/06/2019	Despesa	DESPESAS COM PROJETO PROGRAMA DE EDUCAÇÃO TUTORIAL (PET) 2019.	R$ 1.600,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
24/06/2019	Despesa	DESPESAS COM PROJETO PROGRAMA DE EDUCAÇÃO TUTORIAL (PET) 2019.	R$ 1.600,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
26/06/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 3.200,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
15/08/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 3.200,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
12/09/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 3.200,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET
14/10/2019	Despesa	PAGAMENTO DE BOLSA PET 2019.	R$ 6.400,00	Ensino	Auxílio financeiro a estudantes	Programa de Educacao Tutorial PET

Fonte: Produzida pelo autor.

7.3.3 Gráfico de Linha

A Figura 20 demonstra a implementação da ferramenta Gráfico de Linha para exibição da evolução de gastos com categorias, destacando-se nesta figura especificamente a categoria *"Auxílios financeiros"*. Nota-se também a opção de informação complementar, ao se posicionar o cursor em pontos de interesse, de modo que a ferramenta exibe informação do valor exato da categoria no ponto de análise.

Figura 20 – Série temporal com a evolução de gastos com *"Auxílios financeiros"*.

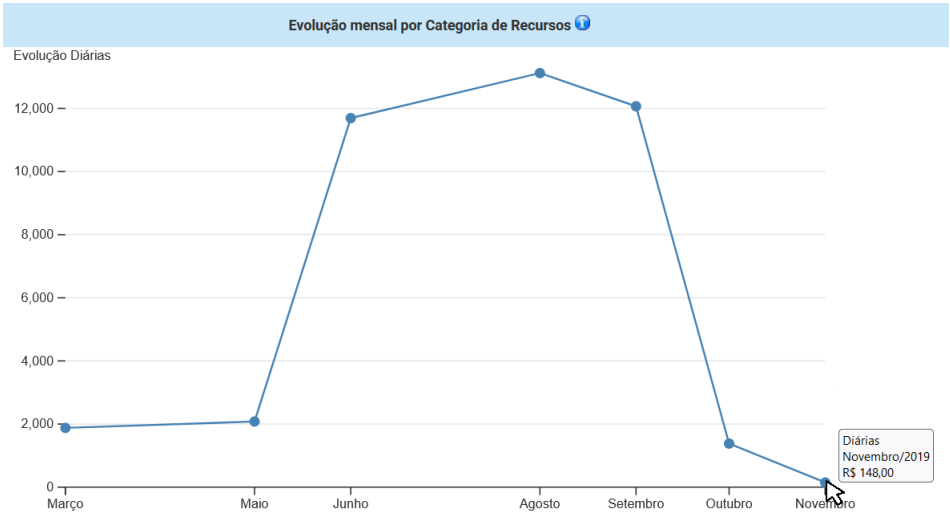


Fonte: Produzida pelo autor.

A Figura 21 apresenta a evolução de gastos com a categoria *"Diárias"*. A análise do gráfico exposto na figura, permite por exemplo, identificar que os períodos de maior atividade externa dos servidores está concentrado no final do primeiro semestre letivo e início do segundo semestre. Nota-se também a ausência de gastos nesta categoria nos meses de janeiro, fevereiro, julho e dezembro, os quais são tradicionalmente relacionados com períodos de férias na universidade.

Os pontos destacados com o cursor nas Figuras 20 e 21, são exemplos de *outliers* para estas categorias.

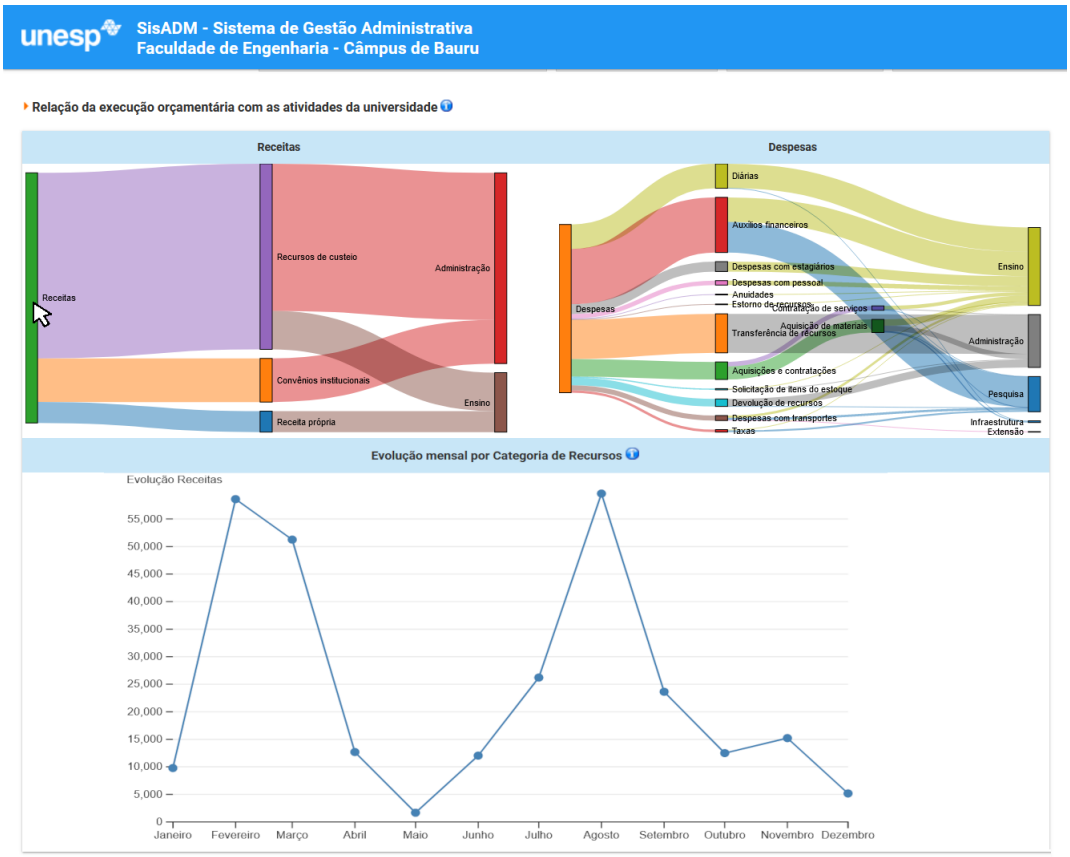
Figura 21 – Série temporal com a evolução de gastos com *"Diárias"*.



Fonte: Produzida pelo autor.

A partir da funcionalidade de acompanhamento da execução orçamentária do ano atual apresentada na Figura 15, é possível acessar a ferramenta Gráfico de Linha para visualizar a representação série temporal de interesse. A Figura 22 demonstra este comportamento ao clicar-se sobre o retângulo que representa a categoria *"Receitas"*.

Figura 22 – Integração da evolução temporal no acompanhamento da execução orçamentária.



Fonte: Produzida pelo autor.

7.4 Implantação da Solução

A solução apresentada neste capítulo foi implantada na Faculdade de Engenharia no segundo semestre de 2021, partindo da integração da solução ao sistema institucional SisADM, concretizando-se com a disponibilização do sistema no ambiente de produção e apresentação da solução para a comunidade da FEB.

O lançamento da solução proposta ocorreu formalmente na 365º reunião da Congregação da Faculdade de Engenharia – Campus de Bauru, realizada ao vigésimo dia do mês de setembro de dois mil e vinte e um, às oito horas e trinta minutos, presidida pelo Prof. Dr. Luttgardes de Oliveira Neto. Após a divulgação oficial da solução proposta, foram realizadas sessões de treinamento com os gestores dos centros de custo estabelecidos para na Faculdade.

Por fim, a administração da Faculdade de Engenharia estabeleceu as diretrizes de acesso e utilização da solução desenvolvida.

7.5 Avaliação da Solução Proposta

Conforme descrito na Seção 1.3, a entrevista aberta foi o método de avaliação escolhido para obter-se respostas qualitativas à solução apresentada.

7.5.1 *Feedback* de Usuários

Esta seção apresenta transcrições de manifestações realizadas por usuários gestores de centros de custo, destacando-se algumas falas de chefes de departamentos:

> *"É bem legal essa ferramenta, vai ajudar muito a gente, porque tem alguns editais e algumas coisas que estão perguntando como a gente está gastando os recursos, no que a gente está aplicando e tal... Ela é mais amigável! Pelos relatórios de saldos, faz quatro anos que estou na chefia e até hoje não sei lidar com eles. Aqui, se você quer ver em bolinha, está lá! Quer ver em números, é só clicar em cima!"* (CDASPP, Sessão de Treinamento II, em 23/11/2021)

> *"Eu gostei bastante do sistema, é bem interessante, é bastante visual, se precisar de mais informações é só passar o mouse em cima que você consegue ver. Temos que dar uma treinada para pegar o jeito, mas tenho certeza que vai ajudar bastante na gestão do departamento."* (CDSDRO, Sessão de Treinamento V, em 25/11/2021)

> *"Até então o chefe de departamento não tinha autonomia nenhuma para realizar essa classificação, tanto que esses gráficos estão aí porque alguém classificou previamente né? Eu nunca tinha tido a oportunidade de alterar a finalidade de uma compra ou um serviço!"* (CDSDRO, Sessão de Treinamento V, em 25/11/2021)

> *"Eu falava bastante com o diretor administrativo, que em geral os chefes de departamento não são economistas, é o pessoal que está ali no momento. Então, no cenário meio geral, o chefe de departamento sempre tinha dificuldade naquele sistema anterior para poder tirar dados de custeio e assim por diante. As vezes era um pouco difícil de controlar as coisas, tinha muita informação e essas informações ficavam um pouco dispersas. A gente tinha que ficar ligando para o diretor administrativo e ficar perguntando onde ficava isso e onde ficava aquilo,*

> *etc. Da minha parte, eu gostei bastante, agora é praticar!"* (CDSDRO, Sessão de Treinamento V, em 25/11/2021)

Destacando-se também o registro realizado por diretor de área: *"Parabéns a todas as pessoas envolvidas nas fases deste projeto, desde sua iniciativa, planejamento, execução e materialização. Excelente ferramenta de apoio para gestão."* (informação verbal)[1].

7.5.2 Dúvidas e Problemas Relatados

Durante as sessões de apresentação e treinamento, um total de dez dúvidas e três problemas foram registradas pelos usuários. Sendo que a maioria das dúvidas manifestadas não estavam relacionadas com as ferramentas de visualização apresentadas.

Foram registradas seis dúvidas sobre a classificação dos registros financeiros e contábeis. Das quais uma dúvida tinha relação ao estabelecimento dos conjuntos de categorias e finalidades passiveis de serem atribuídos aos registros, se estes conjuntos estavam fixados ou se cada gestor poderia criar seu próprio conjunto de classes. O usuário foi informado que os conjuntos eram fixos e foram estabelecidos criteriosamente por membros da área administrativa. Duas dúvidas foram sobre a classificação atribuída à um registro de despesa, os usuários questionaram qual a origem da informação de que aquelas despesas tinham a finalidade de ensino. Forneceu-se então explicação sobre a classificação de registros e como proceder para alterar a classificação de um registro do *Data Mart*. E as últimas três dúvidas neste tópico, foram relacionadas com o procedimento operacional para reclassificação de registros, sendo solucionadas pontualmente e fornecendo-se explicações sobre os dados utilizados nos gráficos serem separados dos dados principais do sistema SisADM, de modo que eles poderiam mexer livremente sem medo de editar ou apagar informações do sistema institucional.

Duas dúvidas estavam relacionadas com a organização hierárquica dos centros de custo e o acesso que os usuários teriam para visualizar dados dos seus próprios centros de custo ou de terceiros. As dúvidas foram respondidas com o fornecimento de explicações sobre o conceito de centro de custos (descrito na Figura 7) e detalhamento de permissões de acesso que a administração da faculdade concedeu aos usuários gestores. Ainda relacionado com os centros de custo, houve apontamento de um problema nos dados exibidos para o centro de custo com sigla STA, que estava exibindo em conjunto os dados de outro centro de custo com a sigla STAEPE, foi aplicada uma correção na filtragem de dados para sanar o problema.

Para a visualização em *dashboard*, um usuário questionou sobre a possibilidade de filtrar dados exibidos, para coletar informações específicas sobre receitas, custeio e convênios. A dúvida foi sanada com a explicação detalhada do funcionamento das opções de filtragem de dados apresentadas na Figura 13. Ainda nesta visualização, outro usuário pontuou que seria melhor corrigir o gráfico de rosca quando a informação da categoria fosse muito pequena ao ponto de nem caber o título da categoria, para que nenhuma informação que conste relacionada nos registros deixe de ser visualizada nos gráficos. A ocorrência do problema apontado pode ser observado na Figura 10, sendo a navegação em profundidade na categoria de interesse, a solução oferecida para esta situação.

[1] Manifestação de Diretor de Colégio Técnico, via *chat*, na Sessão de Apresentação II, em 14/09/2021.

O Diagrama de Sankey apresentado na página inicial dos centros de custo, proporcionou o registro de uma dúvida e a indicação de um problema. A dúvida do usuário estava relacionada a possibilidade de movimentar os itens apresentados no diagrama, que após a movimentação o usurário gostaria de voltar à apresentação original. Foi explicado ao usuário que as mudanças na visualização atual não foram salvas, portanto seria possível retornar à visualização original recarregando a página no navegador. O comentário considerado como apontamento de problema, discorre sobre o espaço disponível para os itens apresentados no diagrama: ‘*“só uma coisa, seria melhor ter mais espaço para poder mexer aqui nos registros, porque o meu aqui tem muito registro”* (CDASPP, Sessão de Treinamento II, em 23/11/2021).

7.5.3 *Feedback* de Cliente

Do ponto de vista da alta administração da faculdade, a solução oferecida atingiu todos os objetivos relacionados com o plano de gestão da diretoria, proporcionando transparência e facilidade de acesso às informações aos gestores, dando subsídios para estes realizarem o planejamento da execução orçamentária para o próximo exercício no ano de 2022 (informação verbal)[2]. Além de atingir os objetivos do plano de gestão, as ferramentas desenvolvidas superaram muito as expectativas da área administrativa, os gráficos e informações disponibilizadas são muito mais do que esperava-se ser possível (informação verbal)[3].

A solução implantada ainda foi mencionada no Relatório de Gestão da Diretoria da Faculdade de Engenharia (2017-2021). Oliveira Neto, Ulson e Barci (2021, pág. 43-46) descrevem que o sistema elaborado pela informática em conjunto com a administração, permite que informações e relatórios da gestão administrativa sejam melhor inseridos, manipulados e gerenciados. Possibilitando o acesso à essas informações de maneira atualizada e transparente, nas condições exigidas pelo órgão controlador externo e atendendo ao anseio da comunidade interna da Faculdade. Os autores destacam que algumas planilhas eram ferramentas de trabalho exclusivas da diretoria administrativa e da seção de finanças, mas agora o sistema surge como uma interface amigável e atualizada disponível para todos.

7.5.4 Resultado da Avaliação da Solução Proposta

Considerando que os problemas apontados pelos usuários não são considerados problemas graves, pois não impedem que os usuários realizem as ações de interesse nas ferramentas de visualização. Considerando também as manifestações espontâneas dos chefes de departamento CDASPP e CDSDRO, corroboradas pelo registro realizado por Oliveira Neto, Ulson e Barci (2021). Admite-se então, que a solução apresentada demonstrou-se amigável e eficiente para gestores de centros de custo, possibilitando a autonomia na gestão e facilitando a tomada de decisões.

[2] Discurso do Diretor da Faculdade de Engenharia, na abertura da Sessão de Apresentação II, em 14/09/2021.
[3] Fala do Diretor Técnico da Divisão Administrativa, na Sessão de Apresentação I, em 01/07/2021.

7.6 Considerações Finais sobre o Protótipo da Solução

Este capítulo dedicou-se à apresentação detalhada da solução proposta para o problema deste trabalho, descrevendo o cenário de implantação, ferramentas utilizadas, o processo de implantação e a avaliação da solução.

8 Considerações Finais

Este capítulo dedica-se à apresentação de possibilidades para trabalhos futuros e discorre sobre limitações deste trabalho.

8.1 Trabalhos Futuros

A possibilidade de trabalho futuro mais iminente, seria a implantação de forma institucional da solução desenvolvida neste trabalho, cobrindo-se toda estrutura da universidade e analisando-se os impactos da implantação.

Também destacam-se possibilidades de trabalhos futuros com assuntos pontuais, como a anonimização automática de dados pessoais contidos em campos descritivos de registros financeiros e contábeis; e a possibilidade de divulgação pública da solução desenvolvida.

8.2 Limitações

Como qualquer estudo empírico, este também apresenta limitações.

Quanto à classificação de texto, por exemplo, o conjunto de dados experimental foi manualmente classificado, o que poderia implicar em erro humano na classificação. Da mesma forma que a criação das classes e seus conjuntos de palavras-chave também foi realizado por humanos. Para mitigar estes possíveis erros humanos, a classificação do conjunto de dados experimental foi realizada por um grupo de pessoas e validada por terceiros, que não participaram do processo de classificação. Enquanto as classes e palavras-chave relacionadas foram estabelecidas de acordo com cada problema de classificação, considerando informações institucionais e registros históricos dos últimos cinco anos anteriores aos registros do *dataset* experimental.

A respeito das ferramentas de visualização, existem inúmeras outras ferramentas que poderiam ser utilizadas no projeto, tendo ainda um espaço de design gigantesco para criação de novas ferramentas. Desta forma, seria inviável a exploração de todas as possibilidades. Para minimizar o erro na escolha das ferramentas trabalhadas, optou-se por realizar a escolha de ferramentas pautando-se no referencial teórico utilizado. No qual especialistas da área de visualização da informação realizam a indicação de tipos adequados de ferramentas conforme a necessidade a ser atendida.

9 Conclusão

O principal problema abordado neste trabalho foi a elucidação de registros públicos financeiros e contábeis. Neste sentido, este trabalho apresentou um estudo de caso sobre a construção e integração de solução de visualização da informação, com objetivo de elucidar a execução orçamentária da Unesp, transformando a complexa prestação de contas da execução orçamentária, em informações visuais e de fácil entendimento.

Partindo-se da questão de como a Informática poderia contribuir na elucidação de registros públicos contábeis, considerados genéricos e pouco informativos ao público de interesse, realizou-se o pré-processamento de dados dos registros financeiros e contábeis da universidade, aplicando-se a classificação automática de texto. Desta forma, foi possível atribuir uma finalidade qualitativa ao empenho de recursos públicos, demonstrando-se para que o recurso financeiro está sendo empregado. Atribuindo-se também uma categoria discriminativa, eliminando generalizações e estabelecendo efetivamente em que o recurso foi empregado.

A avaliação da solução de visualização da informação implantada na Faculdade de Engenharia – Campus de Bauru (FEB), demonstrou que a solução desenvolvida atende tanto as necessidades de informações gerenciais para os gestores de centros de custo, quanto a possibilidade do acompanhamento da execução orçamentária na universidade. Sendo esta solução considerada muito mais amigável na perspetiva do usuário humano, ao transmitir de forma simples as informações de como e com qual finalidade estão sendo utilizados os recursos financeiros da universidade. Conclui-se então, que a solução desenvolvida neste trabalho, é uma solução de análise visual amigável e eficiente para prestação de contas e exploração da execução orçamentária da Unesp.

Demonstram-se verdadeiras todas as quatro hipóteses inicialmente estabelecidas no estudo. Considerando os resultados da avaliação da solução, que demonstram que a solução é eficaz na elucidação da execução orçamentária da universidade, considerando também que a transparência requer uma linguagem acessível e de fácil compreensão para qualquer cidadão (LIMA, 2017; RODRIGUES, 2011). Confirma-se então a Hipótese 1, sobre ferramentas de visualização da informação criadas com informações padronizadas por ferramenta de classificação, serem eficazes para transparência pública e eficientes para elucidação da execução orçamentária. Considerando as técnicas e métodos aplicados no pré-processamento dos dados utilizados nas ferramentas de visualização, confirma-se a Hipótese 2, sobre técnicas e métodos da área de aprendizado de máquina poderem ser integradas ao processo de construção de ferramentas de visualização da informação. Confirmam-se também as Hipóteses 3 e 4, sobre códigos de classificações contábeis poderem ser utilizados como espaços de rótulos estruturados e estes serem utilizados como facilitadores na tarefa de classificação, considerando-se a integração realizada entre o classificador de texto proposto neste trabalho e o sistema institucional SisADM, de modo que o classificador tomou proveito de alguns códigos estruturados disponíveis na legislação.

De forma secundária, abordou-se neste trabalho o estabelecimento de um processo

para construção e integração de ferramentas de visualização em um sistema institucional em funcionamento. O estudo realizado indica que é possível estabelecer tal processo, sendo este constituído pelas atividades elencadas na construção da solução, descritas na fase de execução do estudo (Item 2.4.3):

a) Aquisição de Dados;

b) Armazenamento de Dados;

c) Padronização de Dados;

d) Categorização de Registros;

e) Criação de Protótipos de Ferramentas de Visualização;

f) Criação de Protótipo da Solução;

g) Integração da Solução;

h) Divulgação da Solução;

i) Avaliação da Solução.

Por fim, destaca-se que a realização desta pesquisa demonstrou-se fundamental para compreensão da visão sociotécnica de sistemas de informação, ao desenvolver-se um estudo de caso para resolver um problema do mundo real, realizando aprofundado estudo, com aplicação consistente de métodos e práticas interdisciplinares na solução do problema.

Referências

ACM. *ACM Computing Classification System*. 2012. Disponível em: <https://dl.acm.org/ccs>. Acesso em: 01 dez. 2020. Citado na página 151.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Machine learning*, Springer, v. 6, n. 1, p. 37–66, 1991. Citado na página 67.

AMANI, F. A.; FADLALLA, A. M. Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, v. 24, p. 32–58, 2017. ISSN 1467-0895. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1467089515300488>. Citado 8 vezes nas páginas 45, 120, 121, 130, 139, 141, 142 e 143.

APE. *Anuário Estatístico 2021*. 2021. Disponível em: <https://ape.unesp.br/anuario/pdf/Anuario_2021.pdf>. Acesso em: 22 set. 2021. Citado 3 vezes nas páginas 27, 28 e 29.

APE. *Proposta Orçamentária 2021*. 2021. Disponível em: <https://www2.unesp.br/Home/ape/orcamento_marco2021_co.pdf>. Acesso em: 22 set. 2021. Citado na página 28.

BEC. *Catálogo de Materiais – Pesquisa Avançada*. 2022. Disponível em: <https://www.bec.sp.gov.br/BEC_Catalogo_ui/CatalogoPesquisaAvancada.aspx?chave=&pesquisa=Y&cod_id=&ds_item=>. Acesso em: 22 dez. 2022. Citado na página 54.

BEC. *Catálogo de Serviços – Pesquisa Avançada*. 2022. Disponível em: <https://www.bec.sp.gov.br/BEC_Servicos_UI/V2/CatalogoPesquisaAvancada_V2.aspx?chave=&pesquisa=Y&cod_id=&ds_item=>. Acesso em: 22 dez. 2022. Citado na página 54.

BI, S. et al. Mining knowledge within categories in global and local fashion for multi-label text classification. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2020. p. 1–8. Citado 5 vezes nas páginas 120, 126, 130, 133 e 140.

BITTENCOURT, M. M.; SILVA, R. M.; ALMEIDA, T. A. Ml-mdltext: An efficient and lightweight multilabel text classifier with incremental learning. *Applied Soft Computing*, v. 96, p. 106699, 2020. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494620306372>. Citado 5 vezes nas páginas 120, 123, 130, 133 e 140.

BORKO, H.; BERNICK, M. Automatic document classification. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 10, n. 2, p. 151–162, 1963. Citado na página 44.

BOSCARIOLI, C.; ARAUJO, R. M. D.; MACIEL, R. S. P. I grandsi-br grand research challenges in information systems in brazil 2016-2026. SBC-Sociedade Brasileira de Computação, 2017. Citado 2 vezes nas páginas 21 e 22.

BOSTOCK, M. *D3 – Data-Driven Documents*. d3js, 2021. Disponível em: <https://d3js.org/>. Acesso em: 11 abr. 2023. Citado na página 76.

CASTILLO, P. N. *Mastering D3.js*. Birmingham, England: Packt Publishing Ltd, 2014. ISBN 978-1-78328-627-0. Citado na página 43.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Acm New York, NY, USA, v. 2, n. 3, p. 1–27, 2011. Citado na página 68.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado 2 vezes nas páginas 75 e 122.

CHENG, M.; NAZARIAN, S.; BOGDAN, P. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In: *Proceedings of The Web Conference 2020*. New York, NY, USA: Association for Computing Machinery, 2020. (WWW '20), p. 2892–2898. ISBN 9781450370233. Disponível em: <https://doi-org.ez87.periodicos.capes.gov.br/10.1145/3366423.3380054>. Citado 7 vezes nas páginas 50, 120, 123, 127, 130, 140 e 149.

CLEARY, J. G.; TRIGG, L. E. K*: An instance-based learner using an entropic distance measure. In: *Machine Learning Proceedings 1995*. [S.l.]: Elsevier, 1995. p. 108–114. Citado na página 67.

DA SILVA RODRIGUES, H. L.; BREGA, J. R. F. Visualização da informação como ferramenta de apoio ao tratamento de dados empresariais. In: *Colloquium Exactarum. ISSN: 2178-8332*. [s.n.], 2017. v. 9, n. 2, p. 114–130. Disponível em: <https://revistas.unoeste.br/index.php/ce/article/view/1656>. Citado na página 38.

DAS, A. S.; MEHTA, S.; SUBRAMANIAM, L. Annofin - a hybrid algorithm to annotate financial text. *Expert Systems with Applications*, v. 88, p. 270–275, 2017. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417417304852>. Citado 5 vezes nas páginas 120, 123, 130, 132 e 139.

DERMEVAL, D.; COELHO, J.; BITTENCOURT, I. I. Mapeamento sistemático e revisão sistemática da literatura em informática na educação. *JAQUES, Patrícia Augustin; PIMENTEL, Mariano; SIQUEIRA; Sean; BITTENCOURT, Ig.(Org.) Metodologia de Pesquisa em Informática na Educação: Abordagem Quantitativa de Pesquisa. Porto Alegre: SBC*, 2019. Citado 4 vezes nas páginas 109, 110, 112 e 116.

EIBE, F.; HALL, M. A.; WITTEN, I. H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. 4. ed. [S.l.]: Morgan Kaufmann, 2016. Citado 2 vezes nas páginas 24 e 68.

ELMAN, J. L. Finding structure in time. *Cognitive science*, Wiley Online Library, v. 14, n. 2, p. 179–211, 1990. Citado 2 vezes nas páginas 156 e 157.

ESHIMA, S. *Supplementary Appendix for "Keyword Assisted Topic Models"*. Tese (Thesis) — Department of Political Science, Massachusetts Institute of Technology, 2020. Citado na página 140.

ESHIMA, S.; IMAI, K.; SASAKI, T. Keyword assisted topic models. *arXiv preprint arXiv:2004.05964*, 2020. Citado na página 140.

FEW, S. *Information dashboard design: The effective visual communication of data*. Sebastopol, CA: O'Reilly Media, Inc., 2006. ISBN 0596100167. Citado na página 43.

FISHER, I. E.; GARNSEY, M. R.; HUGHES, M. E. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, v. 23, n. 3, p. 157–214, 2016. ISSN 1055-615X. Citado 13 vezes nas páginas 50, 119, 120, 121, 130, 131, 132, 133, 139, 143, 149, 152 e 159.

FRANK, E.; HALL, M.; PFAHRINGER, B. Locally weighted naive bayes. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2003*. [S.l.: s.n.], 2003. Citado na página 67.

FREITAS, A.; CARVALHO, A. A tutorial on hierarchical classification with applications in bioinformatics. *Research and trends in data mining technologies and applications*, IGI Global, p. 175–208, 2007. Citado na página 47.

FRIEDBERG, R. M. A learning machine: Part i. *IBM Journal of Research and Development*, v. 2, n. 1, p. 2–13, 1958. Citado na página 152.

GIANNOPOULOU, E.; MITROU, N. Extensive experimental evaluation of self-organizing maps for automatic classification of a multi-class multi-label corpus. *IEEE Access*, v. 6, p. 67385–67403, 2018. ISSN 2169-3536. Citado 7 vezes nas páginas 120, 128, 130, 132, 139, 144 e 160.

GIL, A. C. *Como elaborar projetos de pesquisa*. [S.l.]: Atlas São Paulo, 2002. v. 4. Citado na página 23.

GOMES, D. d. S. Inteligência artificial: conceitos e aplicações. *Olhar Científico. v1*, n. 2, p. 234–246, 2010. Citado na página 151.

GONG, C.; SHI, K.; NIU, Z. Hierarchical text-label integrated attention network for document classification. In: *Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference*. New York, NY, USA: Association for Computing Machinery, 2019. (HPCCT 2019), p. 254–260. ISBN 9781450371858. Disponível em: <https://doi-org.ez87.periodicos.capes.gov.br/10.1145/3341069.3342987>. Citado 4 vezes nas páginas 120, 127, 130 e 140.

GOOGLE (Ed.). *Sankey Diagram*. 2022. Disponível em: <https://developers.google.com/chart/interactive/docs/gallery/sankey#overview>. Acesso em: 24 mar. 2023. Citado na página 42.

GUO, L.; SHI, F.; TU, J. Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, v. 2, n. 3, p. 153–170, 2016. ISSN 2405-9188. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2405918816300496>. Citado 12 vezes nas páginas 44, 117, 120, 122, 123, 130, 131, 139, 141, 142, 143 e 148.

HALL, P. A. V.; DOWLING, G. R. Approximate string matching. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 12, n. 4, p. 381–402, 12 1980. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/356827.356830>. Citado na página 60.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 141.

HINGMIRE, S. et al. Document classification by topic labeling. In: . New York, NY, USA: Association for Computing Machinery, 2013. (SIGIR '13), p. 877–880. ISBN 9781450320344. Disponível em: <https://doi-org.ez87.periodicos.capes.gov.br/10.1145/2484028.2484140>. Citado na página 45.

HIRSCHBERG, D. S. Algorithms for the longest common subsequence problem. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 24, n. 4, p. 664–675, 1977. Citado na página 60.

HUANG, K.-W. Exploring the information contents of risk factors in sec form 10-k: A multi-label text classification application. *Available at SSRN 1784527*, 2010. Citado 6 vezes nas páginas 120, 123, 130, 132, 139 e 144.

HUANG, K.-W.; LI, Z. A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Transactions on Management Information Systems (TMIS)*, v. 2, n. 3, p. 1–19, 2011. ISSN 2158-656X. Citado 7 vezes nas páginas 117, 120, 124, 130, 132, 139 e 144.

HUANG, W. et al. Hierarchical multi-label text classification: An attention-based recurrent network approach. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2019. (CIKM '19), p. 1051–1060. ISBN 9781450369763. Disponível em:

<https://doi-org.ez87.periodicos.capes.gov.br/10.1145/3357384.3357885>. Citado 6 vezes nas páginas 120, 125, 127, 130, 132 e 140.

IBRAHIM, D. An overview of soft computing. *Procedia Computer Science*, v. 102, p. 34–38, 2016. ISSN 1877-0509. 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050916325467>. Citado na página 45.

IBRAHIM, M. A. et al. Ghs-net a generic hybridized shallow neural network for multi-label biomedical text classification. *Journal of Biomedical Informatics*, v. 116, p. 103699, 2021. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046421000289>. Citado 6 vezes nas páginas 45, 120, 125, 130, 133 e 140.

IEEE. *IEEE Taxonomy: A Subset Hierarchical Display of IEEE Thesaurus Terms*. 2017. Disponível em: <https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/taxonomy_v101.pdf>. Acesso em: 01 dez. 2020. Citado na página 151.

INGRAM, R. W.; FRAZIER, K. B. Environmental performance and corporate disclosure. *Journal of accounting research*, JSTOR, p. 614–622, 1980. Citado 3 vezes nas páginas 45, 131 e 143.

IRVING, R. W.; FRASER, C. B. Two algorithms for the longest common subsequence of three (or more) strings. In: SPRINGER. *Annual Symposium on Combinatorial Pattern Matching*. [S.l.], 1992. p. 214–229. Citado na página 60.

JEFFCOCK, P. *What's the Difference Between AI, Machine Learning, and Deep Learning?* 2018. Disponível em: <https://blogs.oracle.com/bigdata/difference-ai-machine-learning-deep-learning>. Acesso em: 01 dez. 2020. Citado na página 151.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. [S.l.: s.n.], 1995. p. 338–345. Citado na página 68.

KIRK, A. *Data Visualization: a successful design process*. [S.l.]: Packt publishing LTD, 2012. Citado 4 vezes nas páginas 24, 41, 42 e 43.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. *Technical Report EBSE 2007-001*, Keele University and Durham University Joint Report, 2007. Citado 2 vezes nas páginas 23 e 109.

KOTU, V.; DESHPANDE, B. Chapter 4 - classification. In: KOTU, V.; DESHPANDE, B. (Ed.). *Data Science (Second Edition)*. Second edition. Morgan Kaufmann, 2019. p. 65–163. ISBN 978-0-12-814761-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128147610000046>. Citado na página 44.

KUMAR, B. S.; RAVI, V. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, v. 114, p. 128–147, 2016. ISSN 0950-7051. Citado 20 vezes nas páginas 44, 47, 48, 50, 119, 120, 121, 130, 133, 139, 142, 143, 146, 147, 149, 150, 152, 153, 154 e 155.

LI, J. et al. Multi-label maximum entropy model for social emotion classification over short text. *Neurocomputing*, v. 210, p. 247–256, 2016. ISSN 0925-2312. SI:Behavior Analysis In SN. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231216305987>. Citado 5 vezes nas páginas 120, 123, 130, 132 e 140.

LIMA, M. Q. C. ST 10 Entre transparência e opacidade: o papel da informação no combate a políticas urbanas excludentes. *Anais ENANPUR*, v. 17, n. 1, 2017. Citado na página 93.

LIU, H. et al. Multi-label text classification via joint learning from label embedding and label correlation. *Neurocomputing*, v. 460, p. 385–398, 2021. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231221010754>. Citado 5 vezes nas páginas 120, 125, 127, 130 e 140.

LIU, J. et al. Deep learning for extreme multi-label text classification. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2017. (SIGIR '17), p. 115–124. ISBN 9781450350228. Disponível em: <https://doi-org.ez87.periodicos.capes.gov.br/10.1145/3077136.3080834>. Citado 6 vezes nas páginas 75, 120, 125, 130, 132 e 140.

MA, Y. et al. Hybrid embedding-based text representation for hierarchical multi-label text classification. *Expert Systems with Applications*, v. 187, p. 115905, 2022. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417421012604>. Citado 7 vezes nas páginas 45, 66, 120, 126, 130, 133 e 140.

MAALEJ, W. et al. On the automatic classification of app reviews. *Requirements Engineering*, Springer, v. 21, n. 3, p. 311–331, 2016. Citado na página 44.

MALTOUDOGLOU, L. et al. Well-calibrated confidence measures for multi-label text classification with a large number of labels. *Pattern Recognition*, v. 122, p. 108271, 2022. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320321004519>. Citado 8 vezes nas páginas 120, 124, 126, 127, 130, 133, 140 e 159.

MARON, M. E. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 8, n. 3, p. 404–417, 1961. Citado na página 44.

MEDEIROS, J. F. *Tag The Web USING WIKIPEDIA CATEGORIES TO AUTOMATICALLY CATEGORIZE TEXT-BASED RESOURCES ON THE WEB*. Tese (Thesis) — Universidade Federal do Estado do Rio de Janeiro, 2018. Citado 10 vezes nas páginas 48, 120, 125, 130, 131, 133, 139, 143, 146 e 147.

METZ, J. *Abordagens para aprendizado semissupervisionado multirrótulo e hierárquico*. Tese (Thesis) — Instituto de Ciências Matemáticas e de Computação (ICMC-USP), São Carlos, São Paulo, Brasil, 2011. Citado 9 vezes nas páginas 47, 120, 125, 130, 139, 144, 145, 146 e 150.

MICHIE, D. "memo" functions and machine learning. *Nature*, Nature Publishing Group, v. 218, n. 5136, p. 19–22, 1968. Citado na página 152.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119. Citado 2 vezes nas páginas 158 e 159.

MIROŃCZUK, M. M.; PROTASIEWICZ, J. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, v. 106, p. 36–54, 2018. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S095741741830215X>. Citado 16 vezes nas páginas 45, 47, 48, 114, 119, 120, 121, 130, 132, 139, 144, 146, 147, 148, 150 e 152.

MUNZNER, T. *Visualization analysis and design*. [S.l.]: CRC press, 2014. ISBN 9781466508910. Citado 6 vezes nas páginas 20, 37, 38, 39, 40 e 43.

MUSTAFI, D.; MUSTAFI, A.; SAHOO, G. A novel approach to text clustering using genetic algorithm based on the nearest neighbour heuristic. *International Journal of Computers and Applications*, p. 1–13, 2020. ISSN 1206-212X. Citado 6 vezes nas páginas 120, 128, 130, 139, 148 e 160.

NARCISO, E. N.; NUNES, F. L.; DELAMARO, M. E. Seleção de casos de teste utilizando conceitos de variabilidade: Uma revisão sistemática. *Anais do VIII Simp. Bras. de Sistemas de Informação, São Paulo–SP, Brasil*, p. 115–125, 2011. Citado 2 vezes nas páginas 110 e 112.

Oliveira Neto, L. d.; ULSON, J. A. C.; BARCI, J. d. A. *Relatório de Gestão da Diretoria 2017-2021*. Faculdade de Engenharia, FEB, 2021. Disponível em: <https://drive.google.com/file/d/1w-Yj6LZyCFCy9HzqRaBo0CHk4DkGjDZN/view>. Acesso em: 21 mar. 2022. Citado na página 90.

ORACLE (Ed.). *What is a Data Mart?* 2023. Disponível em: <https://www.oracle.com/autonomous-database/what-is-data-mart/>. Acesso em: 20 mar. 2023. Citado na página 35.

PIKIES, M.; ALI, J. String similarity algorithms for a ticket classification system. In: IEEE. *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*. [S.l.], 2019. p. 36–41. Citado 2 vezes nas páginas 45 e 60.

PINK, G. A. *Slot filling*. Tese (Thesis) — The University of Sydney, Austrália, 2017. Citado 4 vezes nas páginas 120, 128, 130 e 139.

PRADHAN, N.; GYANCHANDANI, M.; WADHVANI, R. A review on text similarity technique used in ir and its application. *International Journal of Computer Applications*, Foundation of Computer Science, v. 120, n. 9, p. 29–34, 2015. Citado na página 60.

PRAKOSO, D. W.; ABDI, A.; AMRIT, C. Short text similarity measurement methods: a review. *Soft Computing*, v. 25, n. 6, p. 4699–4723, 3 2021. ISSN 1433-7479. Disponível em: <https://doi.org/10.1007/s00500-020-05479-2>. Citado na página 60.

QUINLAN, J. R. Program for machine learning. *C4.5*, Morgan Kaufmann Pub, 1993. Citado na página 68.

RANIERI, N. B. S. Trinta anos de autonomia universitÁria: Resultados diversos, efeitos contraditÓrios. *Educação & Sociedade [online]*, v. 39, n. 145, p. 946–961, 2018. ISSN 1678-4626. Disponível em: <https://doi.org/10.1590/ES0101-73302018205173>. Citado na página 28.

RAZA, M. et al. A comparative analysis of machine learning models for quality pillar assessment of saas services by multi-class text classification of users' reviews. *Future Generation Computer Systems*, v. 101, p. 341–371, 2019. ISSN 0167-739X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X19300196>. Citado 15 vezes nas páginas 44, 48, 68, 120, 122, 123, 130, 132, 139, 143, 144, 146, 148, 153 e 154.

READ, J. et al. Classifier chains for multi-label classification. *Machine learning*, Springer, v. 85, p. 333–359, 2011. Citado na página 46.

RIOS, A.; KAVULURU, R. Few-shot and zero-shot multi-label learning for structured label spaces. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, v. 2018, p. 3132–3142, 2018. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6375489/>. Citado 7 vezes nas páginas 120, 123, 130, 133, 139, 144 e 158.

ROBINS, M. *The Difference Between Artificial Intelligence, Machine Learning and Deep Learning*. 2020. Disponível em: <https://www.intel.com.au/content/www/au/en/artificial-intelligence/posts/difference-between-ai-machine-learning-deep-learning.html>. Acesso em: 01 dez. 2020. Citado na página 155.

RODRIGUES, S. L. Mídia, informação e transparência construindo a cidadania contra a corrupção no maranhão. In: *Trabalho apresentado no Grupo de Trabalho da II Conferência Sul-Americana e VII Conferência Brasileira de Mídia Cidadã*. [S.l.: s.n.], 2011. Citado na página 93.

SÃO PAULO. Portaria co nº 09, de 17-12-2018 - consolida a classificação da despesa orçamentária por natureza. *Diário Oficial [do] Estado de São Paulo*, São Paulo, SP, v. 128, n. 235, p. 16–20, 2018. Disponível em: <https://www.imprensaoficial.com.br/DO/BuscaDO2001Documento_11_4.aspx?link=%2f2018%2fexecutivo%2520secao%2520i%2fdezembro%2f18%2fpag_0016_a0f3e19776261ade62a0e8b511e714dc.pdf&pagina=16&data=18/12/2018&caderno=Executivo%20I&paginaordenacao=100016>. Acesso em: 23 mar. 2023. Citado na página 106.

SÃO PAULO. *Lei Nº 17.309, de 29 de dezembro de 2020 - Lei Orçamentária 2021*. 2020. Disponível em: <http://orcamento.planejamento.sp.gov.br/download_lei/2021>. Acesso em: 22 set. 2021. Citado 2 vezes nas páginas 20 e 28.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, 1959. Citado na página 152.

SANTOS, V. Batista dos; MERSCHMANN, L. H. d. C. Metalearning applied to multi-label text classification. In: *XVI Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2020. (SBSI'20). ISBN 9781450388733. Disponível em: <https://doi-org.ez87.periodicos.capes.gov.br/10.1145/3411564.3411646>. Citado 11 vezes nas páginas 48, 49, 75, 120, 123, 130, 132, 133, 140, 146 e 148.

SCHRÖDER, K. *Hierarchical Multiclass Topic Modelling with Prior Knowledge*. Tese (Thesis) — Humboldt-Universität zu Berlin, Alemanha, 2018. Citado 12 vezes nas páginas 46, 47, 66, 120, 124, 125, 130, 132, 139, 145, 146 e 149.

SCIENCEDIRECT (Ed.). *Document Classification*. 2023. Disponível em: <https://www.sciencedirect.com/topics/computer-science/document-classification>. Acesso em: 13 jan. 2023. Citado na página 44.

Secretaria da Fazenda e Planejamento. *Portaria CO Nº 09, de 14-12-2018 - Consolida a Classificação da Despesa Orçamentária por Natureza*. 2023. Disponível em: <http://vclippingorcamento.planejamento.sp.gov.br/Vclipping1/images/6/6f/PORTARIA_CO_9_DE_14.12.2018-CONSOLIDA_A_CLASSIFICAĂĞĂČO_DA_DESPESA_ORĂĞAMENTĂĄRIA_POR_NATUREZA_(ATUALIZADA_EM_13.03.2023).pdf>. Acesso em: 23 mar. 2023. Citado 2 vezes nas páginas 20 e 106.

SONG, Y. *Machine Learning for Text Mining: Classification, Retrieval and Recommendation*. Tese (Doutorado) — Pennsylvania State University, EUA, 2009. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.475.6356&rep=rep1&type=pdf>. Citado 10 vezes nas páginas 120, 123, 130, 132, 139, 141, 143, 144, 150 e 154.

SUSMAGA, R. Confusion matrix visualization. In: KŁOPOTEK, M. A.; WIERZCHOŃ, S. T.; TROJANOWSKI, K. (Ed.). *Intelligent Information Processing and Web Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 107–116. ISBN 978-3-540-39985-8. Citado na página 68.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. p. 3104–3112. Citado na página 159.

TAO, Y.; CUI, Z.; WENJUN, Z. A multi-label text classification method based on labels vector fusion. In: *2018 International Conference on Promising Electronic Technologies (ICPET)*. [S.l.: s.n.], 2018. p. 80–85. Citado 8 vezes nas páginas 66, 120, 124, 130, 132, 140, 158 e 159.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, IGI Global, v. 3, n. 3, p. 1–13, 2007. Citado na página 46.

TUFTE, E. R. *The visual display of quantitative information*. 2. ed. Cheshire, CT: Graphics Press, 2001. Citado na página 40.

TUFTE, E. R. *Beautiful evidence*. [S.l.]: Graphics press Cheshire, CT, 2006. v. 1. Citado 2 vezes nas páginas 37 e 42.

UNA-SUS. *Organização Mundial de Saúde declara pandemia do novo Coronavírus*. 2020. Disponível em: <https://www.unasus.gov.br/noticia/organizacao-mundial-de-saude-declara-pandemia-de-coronavirus>. Acesso em: 05 mai. 2021. Citado na página 66.

UNESP. *Perfil*. 2023. Disponível em: <https://www2.unesp.br/portal#!/sobre-a-unesp/perfil/>. Acesso em: 15 mar. 2023. Citado 2 vezes nas páginas 27 e 28.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 1995. Citado na página 154.

WALLACE, S. A.; PAVLENKO, V. Using a document classification task to introduce machine learning. *J. Comput. Sci. Coll.*, Consortium for Computing Sciences in Colleges, Evansville, IN, USA, v. 27, n. 1, p. 188–194, 10 2011. ISSN 1937-4771. Citado na página 44.

WANG, C.; TAN, C. Label-based convolutional neural network for text classification. In: *Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence*. New York, NY, USA: Association for Computing Machinery, 2021. (CCEAI 2021), p. 136–140. ISBN 9781450388870. Disponível em: <https://doi-org.ez87.periodicos.capes.gov.br/10.1145/3448218.3448235>. Citado 4 vezes nas páginas 120, 127, 130 e 140.

WANG, J.; DONG, Y. Measurement of text similarity: A survey. *Information*, v. 11, n. 9, 2020. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/11/9/421>. Citado na página 60.

WARE, C. *Information Visualization: Perception for Design*. [S.l.]: Elsevier, 2012. Citado na página 38.

XIANG, Y.; ZHENG, J. Multi-label emotion classification for imbalanced chinese corpus based on cnn. In: *2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. [S.l.: s.n.], 2018. p. 38–43. Citado 8 vezes nas páginas 75, 120, 122, 130, 132, 140, 158 e 159.

XIAO, Y. et al. History-based attention in seq2seq model for multi-label text classification. *Knowledge-Based Systems*, v. 224, p. 107094, 2021. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705121003579>. Citado 6 vezes nas páginas 120, 124, 126, 130, 133 e 140.

YIN, R. K. *Estudo de Caso-: Planejamento e métodos*. [S.l.]: Bookman editora, 2015. Citado na página 23.

YOUNG, T. et al. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, v. 13, n. 3, p. 55–75, 2018. ISSN 1556-6048. Citado 15 vezes nas páginas 44, 45, 50, 117, 120, 121, 130, 139, 149, 150, 155, 156, 157, 158 e 159.

ZADEH, L. Soft computing and fuzzy logic. *IEEE Software*, v. 11, n. 6, p. 48–56, 1994. Citado na página 45.

ZAYAS, B. et al. Getting ready for data analytics of electric power distribution systems. *International Journal of Computers*, v. 2, 2017. Citado 7 vezes nas páginas 37, 66, 114, 120, 128, 130 e 139.

ZHENG, J.; ZHENG, L. A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification. *IEEE Access*, v. 7, p. 106673–106685, 2019. ISSN 2169-3536. Citado 8 vezes nas páginas 120, 123, 130, 132, 139, 143, 144 e 158.

Apêndices

APÊNDICE A – Projeto de pesquisa

A.1 Título da pesquisa

A utilização de técnicas de visualização da informação na análise de registros públicos contábeis.

A.2 Tema de pesquisa

Criação e integração de ferramentas de Visualização da Informação aos sistemas legados institucionais, utilizando-se conceitos, técnicas ou ferramentas da área de Aprendizado de Máquina, para classificação de registros contábeis e padronização de informações que serão utilizadas nas ferramentas de visualização.

A.3 Objetivos

O objetivo geral do trabalho é realização de um estudo da aplicação conjunta de conceitos, técnicas e ferramentas interdisciplinares, relacionadas com as áreas da Inteligência Artificial, Aprendizado de Máquina e Visualização da Informação. Objetivo a ser alcançado pela realização de um estudo de caso, da implantação e integração de ferramentas visuais, que relacionem cada lançamento contábil com categorias mais amigáveis e menos burocráticas sob a perspetiva humana. Ferramentas que demonstrem também, a relação dos registros contábeis com as atividades-fim e atividades-meio da instituição observada.

A.4 Definição do problema

O principal problema a ser resolvido com este projeto de pesquisa, é a elucidação de registros públicos contábeis. Tendo como foco mais específico, registros que de acordo com a legislação vigente são contabilmente categorizados de forma genérica. Ou seja, a elucidação de registros pouco informativos ao público de interesse.

Um problema secundário a ser abordado, é o estabelecimento de um processo para construção e integração de ferramentas de visualização em um sistema institucional em funcionamento.

A.5 Justificativa e relevância do tema

Frequentemente registros financeiros e contábeis apresentam informações burocráticas e de difícil entendimento. A análise dos registros de uma universidade pública do Estado de São Paulo, motivou o desenvolvimento desta pesquisa. Nestes registros, encontram-se informações relacionadas a categoria e a finalidade de destino do recurso financeiro empregado.

Muitos destes registros são vinculados com categorias contábeis genérica como *"Outros..."*, que fazem parte da classificação contábil padronizada por legislação vigente (SÃO PAULO, 2018) (Secretaria da Fazenda e Planejamento, 2023). Entretanto, estes registros também possuem informações adicionais que podem sugerir uma categoria corretamente discriminativa, ao invés de simplesmente uma categoria genérica.

Além disso, somente alguns registros apresentam codificação contábil relacionada à finalidade do recurso. Mas, da mesma forma que informações adicionais permitem a inferência de categoria específica, é comum a presença de informações que justifiquem a necessidade de determinadas despesas, o que pode sugerir a finalidade do recurso. Estas informações são oriundas de pessoas solicitantes, que alimentam os sistemas institucionais com justificativas redigidas em linguagem natural.

Desta forma, considera-se utilizar a tarefa de Classificação, da área de Mineração de Texto, para padronizar em categorias específicas os registros contábeis, realizando também a identificação e padronização de finalidade do emprego de recurso financeiro. Com objetivo de vincular os registros contábeis com categorias e finalidades padronizadas, que serão trabalhadas por técnicas da área de Visualização da Informação.

Esta pesquisa mostra-se relevante ao desmistificar a execução orçamentária de um órgão público, traduzindo a complexa burocracia incutida nos registros contábeis, em categorias padronizadas e de fácil entendimento, sendo estas também vinculadas com finalidades que refletem as atividades-fim, que formam o tripé base de qualquer universidade (Ensino, Pesquisa e Extensão), além de atividades-meio, como Infraestrutura e Administração.

A.6 Hipóteses

- Hipótese 1 — Ferramentas de visualização da informação, criadas com informações padronizadas por ferramenta de classificação, são eficazes para transparência pública e eficientes para elucidação da execução orçamentária.
- Hipótese 2 — Técnicas e métodos da área de Classificação de Texto, podem ser integradas ao processo de construção de ferramentas de visualização da informação.
- Hipótese 3 — Espaços de rótulos estruturados, podem ser utilizados como facilitadores na tarefa de classificação.
- Hipótese 4 — Códigos e descrições de elementos contábeis, podem ser utilizados como espaços de rótulos estruturados.
- Hipótese 5 — É possível estabelecer um processo para incorporação por sistemas legados, de ferramentas ou técnicas da área de Aprendizado de Máquina.

A.7 Metodologia do projeto de pesquisa

Pesquisa descritiva, realizada através do estudo de caso da aplicação de conceitos e técnicas de Aprendizado de Máquina, como suporte para geração de informações à serem utilizadas em

ferramentas de Visualização da Informação, que deverão ser integradas a um sistema institucional legado.

A.8 Referencial teórico

O referencial teórico será composto pelo resultado da Revisão Sistemática da Literatura (RSL) apresentada no Apêndice B. Complementado por demais pesquisas bibliográficas necessárias para abordagem do tema.

A.9 Resultados esperados

Espera-se como resultado desta pesquisa, a compilação de produção técnica-científica, com foco nos assuntos relacionados com o tema abordado.

◆ Dissertação de mestrado

A produção primária almejada com a pesquisa, será a dissertação de mestrado. No momento de concepção deste projeto de pesquisa, definiu-se que o provável título da dissertação será "A utilização de técnicas de visualização da informação na análise de registros públicos contábeis", que abordará o estudo de caso da implantação de ferramentas e técnicas de visualização da informação, para elucidação de registros contábeis genéricos da universidade estudada.

◆ Artigos científicos

Devido a natureza interdisciplinar da pesquisa, destacam-se as possibilidades de redação de artigos com os possíveis temas:

- Método/Processo para implantação de ferramentas visualização da informação em sistemas legados, apoiados por técnicas ou ferramentas de Aprendizado de Máquina.
- Promoção da transparência pública com a utilização da ferramentas de visualização da informação: Um estudo de caso em uma universidade do Estado de São Paulo.
- Estudo comparativo entre classificador textual empírico e algoritmos de Aprendizado de Máquina.

A.10 Cronograma de atividades

Para melhor visualização do cronograma de atividades da proposta de projeto de pesquisa de mestrado, também para melhor organização desta monografia, dedica-se esta seção à apresentação da Tabela 16, que representa o conjunto de atividades do projeto e seus respectivos períodos previstos para execução.

Tabela 16 – Cronograma de desenvolvimento das atividades do mestrado.

Atividade	2020	Jan 21	Fev 21	Mar 21	Abr 21	Mai 21	Jun 21	Jul 21	Ago 21	Set 21	Out 21	Nov 21	Dez 21	Jan 22	Fev 22	Mar 22	Abr 22	Mai 22	Jun 22	Jul 22	Ago 22	Set 22
Disciplinas comuns																						
Estudos Especiais I																						
Classificador empírico																						
Protótipo visualização																						
Integração classificação																						
Integração visualização																						
Implantação e adequação																						
Redação da qualificação																						
Qualificação																						
Integração aprendizado máquina																						
Experimentos																						
Redação de artigos																						
Redação da dissertação																						
Defesa																						

Fonte: Produzida pelo autor.

A.11 Considerações finais sobre o projeto de pesquisa

Este capítulo discorreu sobre o projeto de pesquisa proposto neste trabalho, apresentando a motivação, estabelecendo hipóteses e objetivos, e, indicando a metodologia e referencial teórico. O cronograma de atividades do projeto de pesquisa será apresentado no capítulo seguinte.

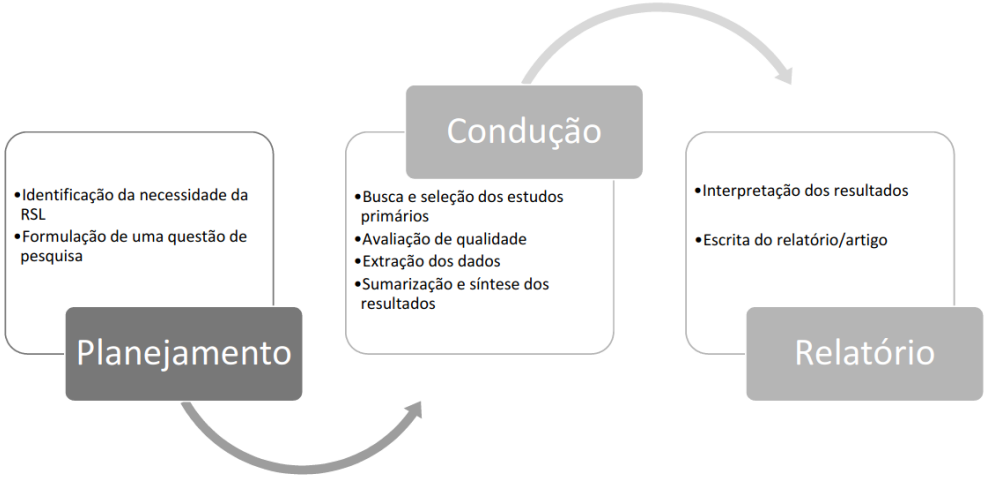
APÊNDICE B – Revisão Sistemática da Literatura

Este capítulo é dedicado à apresentação das estratégias utilizadas na pesquisa bibliográfica, realizada no formato de Revisão Sistemática da Literatura (RSL), com objetivo de identificar contribuições acadêmicas e científicas no estudo das relações entre as áreas de Inteligência Artificial, Aprendizado de Máquina e Visualização da Informação, tendo como foco o processamento de informações textuais relacionadas às áreas de contabilidade e finanças.

O trabalho de Kitchenham e Charters (2007) estabelece diretrizes para um levantamento bibliográfico preciso e consistente. Os autores definem a RSL como uma forma de avaliar e interpretar todas as pesquisas relevantes, que estão disponíveis para uma questão de pesquisa específica, uma área de tópico ou um fenômeno de interesse. As revisões sistemáticas visam apresentar uma avaliação justa de um tópico de pesquisa, através de uma metodologia de pesquisa rigorosa, confiável e auditável.

Dermeval, Coelho e Bittencourt (2019) comentam que o protocolo estabelecido em Kitchenham e Charters (2007), baseia-se em outros protocolos amplamente utilizados em pesquisas baseadas em evidências, afirmando ainda que a RSL é a metodologia mais utilizada em trabalhos sistemáticos de levantamento da literatura na área da Computação. A partir das diretrizes estabelecidas e atividades da RSL descritas em Kitchenham e Charters (2007), os autores organizam e apresentam a revisão sistemática como um processo de três fases: planejamento, condução e relatório. A Figura 23 apresenta a divisão em fases realizada pelos autores, ilustrando também as atividades da RSL relacionada com cada fase.

Figura 23 – Fases da Revisão Sistemática da Literatura.



Fonte: Adaptado de Dermeval, Coelho e Bittencourt (2019).

B.1 Planejamento da RSL

A fase de planejamento da revisão sistemática consiste na identificação da necessidade da revisão e no estabelecimento de diretrizes que guiarão a pesquisa (DERMEVAL; COELHO; BITTENCOURT, 2019). Como resultado desta etapa, espera-se o estabelecimento de um protocolo documentado da revisão sistemática, que pode sofrer alterações durante a execução da revisão (NARCISO; NUNES; DELAMARO, 2011). O protocolo estabelecido para esta RSL encontra-se no Apêndice C.

B.1.1 Motivação da RSL

A motivação inicial para este trabalho partiu da análise de um modelo de visualização criado manualmente, utilizando dados de registros financeiros e contábeis, dados estes que apresentam informações diversificadas e de difícil entendimento. Esta análise preliminar revelou a presença de problemas computacionais relacionados a Classificação de Texto, Processamento de Linguagem Natural e Visualização de Informações. Originando a questão: Como a Informática poderia contribuir para melhoria da compreensão humana sobre esses dados?

B.1.2 Objetivo da RSL

Esta revisão tem por objetivo identificar técnicas, métodos, aplicações, usos e contribuições do Aprendizado de Máquina para área de Visualização da Informação, ao se trabalhar com dados em registros textuais no formato alfanumérico, considerando qual a contribuição ou correlação de aprendizado de máquina com visualização da informação textual, sob a perspectiva humana da relação de Interação Humano-Computador.

B.1.3 Questões de pesquisa

Dermeval, Coelho e Bittencourt (2019) destacam que questão de pesquisa é a atividade mais importante da fase de planejamento de uma revisão sistemática, pois é esta questão que guiará toda a condução da pesquisa. Destacam ainda que a formulação da questão de pesquisa deve ser em função do foco e objetivo da revisão. Desta forma, a partir do objetivo estabelecido se obtêm a principal questão de pesquisa:

- Como a Inteligência Artificial e o Aprendizado de Máquina podem ser utilizados na construção de ferramentas de visualização de informações textuais?

Sendo essa questão principal subdividida em quatro questões objetivas:

1. Quais são os métodos e técnicas de Aprendizado de Máquina que estão inseridos no contexto de processamento de registros textuais?
2. Como é a utilização do Aprendizado de Máquina para processamento de registros textuais?
3. Qual a contribuição do Aprendizado de Máquina, quando utilizado no processamento de textos, para área da Visualização da Informação?

4. Quais são os conceitos e/ou tecnologias que se relacionam com as áreas de Inteligência Artificial e Aprendizado de Máquina, que possam se inter-relacionar com a área da Visualização da Informação?

B.1.4 Termos de busca

Na primeira abordagem, considerando todas as áreas do conhecimento relacionadas (Inteligência Artificial, Aprendizado de Máquina e Visualização da Informação) não foram obtidos resultados que relacionassem os problemas. Decidiu-se então, pela divisão das pesquisas por área do conhecimento: Inteligência Artificial e Visualização da Informação, esta revisão sistemática se concentra na área da Inteligência Artificial.

Após algumas discussões e a realização de pesquisas preliminares, foram selecionados os termos de busca: *"natural language processing"*, *"text classifier"*, *"textual classification"*. Que posteriormente foram complementados por termos relacionados ao objetivo geral deste trabalho: *accounting*, *accountability*, *financial* e *multi-label*. A Tabela 17 exibe histórico de pesquisas exploratórias realizadas no Google Acadêmico para refinamento dos termos de busca.

Tabela 17 – Histórico de pesquisas para refinamento de *string* de busca.

String de busca	Resultados obtidos
"Natural Language Processing" AND (Multilabel OR "Multi-label" OR "Multi Label") AND ((text OR textual) AND (classifier OR classification))	Aproximadamente 11.300
intitle:(Multilabel OR "Multi-label" OR "Multi Label") AND "Natural Language Processing"AND ((text OR textual) AND (classifier OR classification))	Aproximadamente 10.300
(intitle:(text OR textual) AND intitle:(classifier OR classification)) AND intitle:(Multilabel OR "Multi-label" OR "Multi Label") AND "Natural Language Processing"	Aproximadamente 5.130
(intitle:(text OR textual) AND intitle:(classifier OR classification)) AND intitle:(Multilabel OR "Multi-label" OR "Multi Label") AND "Natural Language Processing"AND "Financial Transaction" AND (Accounting OR Accountability)	1 resultado
(intitle:(text OR textual) AND intitle:(classifier OR classification)) AND intitle:(Multilabel OR "Multi-label" OR "Multi Label") AND "Natural Language Processing"AND (Financial OR Accounting OR Accountability)	Aproximadamente 1.230
((text OR textual) AND (classifier OR classification)) AND intitle:(Multilabel OR "Multi-label" OR "Multi Label") AND "Natural Language Processing"AND Financial AND (Accounting OR Accountability)	Aproximadamente 302
(intitle:(text OR textual) AND intitle:(classifier OR classification)) AND intitle:(Multilabel OR "Multi-label" OR "Multi Label") AND "Natural Language Processing"AND Financial AND (Accounting OR Accountability)	Aproximadamente 195

Fonte: Produzida pelo autor.

B.1.5 Bases de busca científica

Inicialmente o Google Acadêmico[1] foi selecionado como base de busca para pesquisa exploratória. Em sequida a *ACM Digital Library*[2] e o *IEEE Xplore*[3] foram bases selecionadas por serem referência na área da Computação. Por fim, a *ScienceDirect*[4] foi adicionada por ser a base com a maior quantidades de trabalhos selecionados durante a pesquisa exploratória.

B.1.6 Critérios de inclusão, exclusão e qualidade

Os critérios de inclusão definidos, buscam a aproximação do estudo aos temas relacionados nos termos de busca, possibilitando a aceitação de artigos primários e secundários. Já os critérios de exclusão, delimitaram a aceitação de artigos quanto ao tópico de pesquisa e a questões práticas como linguagem ou disponibilidade de acesso ao texto integral (DERMEVAL; COELHO; BITTENCOURT, 2019).

A seleção de estudos foi realizada com base em critérios pré-estabelecidos no Protocolo da Revisão Sistemática. O único critério de qualidade estabelecido foi a revisão do trabalho por pares ou aprovação por banca examinadora.

B.2 Condução da RSL

A fase de condução consiste na etapa onde a pesquisa propriamente dita é realizada, iniciando com o processamento e registro das buscas efetuadas, culminando com a extração de dados e síntese dos resultados (NARCISO; NUNES; DELAMARO, 2011).

B.2.1 Buscas da pesquisa exploratória

Inicialmente as buscas foram realizadas no Google Acadêmico, entre os dias 05 e 09 de outubro de 2020. Neste período foram realizadas quatro sessões de busca com objetivos e *strings* distintas. Pelo caráter exploratório, todas as buscas foram realizadas sem delimitação do período de publicação dos trabalhos, mas desconsiderando registros de patentes e citações.

➥ 1ª Iteração de busca

Busca inicial de amplo aspecto, considerando todos os termos de busca estabelecidos. *String* de busca:

(intitle:(text OR textual) AND intitle:(classifier OR classification)) AND intitle:(Multilabel OR “Multi-label” OR “Multi Label”) AND “Natural Language Processing” AND Financial AND (Accounting OR Accountability)

Foram listados 195 resultados, dos quais 19 foram selecionados.

[1] https://scholar.google.com.br
[2] http://dl.acm.org/
[3] http://ieeexplore.ieee.org/
[4] http://www.sciencedirect.com/

➥ 2ª Iteração de busca

Busca por trabalhos com palavras-chave em português. *String* de busca:

((classificador OR classificação) AND (texto OU textual)) AND (Multiclasse OU "Multi-classe" OR "Multi Classe")

Foram listados 31 resultados, nenhum selecionado.

➥ 3ª Iteração de busca

A partir de análise prévia dos trabalhos selecionados nas buscas anteriores, observou-se a necessidade de realizar busca por termo hierárquico. *String* de busca:

"Hierarchical multiclass" classification

Na primeira aplicação desta *string* de busca, foram listados 203 resultados. Considerando a alta quantidade de resultados retornados e motivado pela ausência de resultados selecionados na segunda sessão de busca, foi adicionado o parâmetro de filtragem de apenas páginas em português.

A nova submissão da *string* de busca retornou 3 resultados, sendo 1 selecionado.

➥ 4ª Iteração de busca

A última sessão de busca foi idealizada considerando termos latentes identificados nos estudos previamente selecioandos. *String* de busca:

"Hierarchical Multiclass" AND "Natural Language Processing"

Foram listados 20 resultados, dos quais 1 foi selecionado.

B.2.2 Buscas nas bases selecionadas

Após a realização da pesquisa exploratória, constatou-se a necessidade de adequação da *string* de buscas. O termo *Financial* foi removido, pois está mais relacionado com o mercado financeiro. Também foram adicionados termos que abrangem o conceito multiclasse. Desta forma, estabeleceu-se uma nova *String* de busca para ser submetida às bases científicas:

(intitle:(text OR textual) AND intitle:(classifier OR classification)) AND intitle:(Multilabel OR "Multi-label" OR "Multi Label" OR Multiclass OR "Multi-class" OR "Multi Class") AND "Natural Language Processing" AND (Accounting OR Accountability)

Considerando que na pesquisa exploratória foram selecionados trabalhos das bases científicas escolhidas para busca, considerando também o intuito de encontrar trabalhos mais contemporâneos, delimitou-se o período de busca para os últimos cinco anos nessas bases. Sendo assim, a submissão da *string* de busca às bases foi realizada em setembro de 2021, com filtragem de resultados para publicações a partir de 2016.

❖ *ACM Digital Library*

String de busca específica da base: *[[Publication Title: text] OR [Publication Title: textual]] AND [[Publication Title: classifier] OR [Publication Title: classification]] AND [[Publication Title: multilabel] OR [Publication Title: "multi-label"] OR [Publication Title: multi label]] AND [All: "natural language processing"] AND [[All: accounting] OR [All: accountability]] AND [Publication Date: (01/01/2016 TO *)]*

Foram listados 6 resultados, todos foram selecionados.

❖ *IEEE Xplore*

String de busca específica da base: *("All Metadata":text OR "All Metadata":textual) AND ("All Metadata":classifier OR "All Metadata":classification) AND ("All Metadata":Multilabel OR "All Metadata":"Multi-label" OR "All Metadata":"Multi Label" OR "All Metadata":Multiclass OR "All Metadata":"Multi-class" OR "All Metadata":"Multi Class") AND ("All Metadata":"Natural Language Processing") AND ("All Metadata":Accounting OR "All Metadata":Accountability)*

Foram listados 6 resultados, dos quais 3 foram selecionados.

❖ *ScienceDirect*

Campos de busca específicos da base: *Find articles with these terms* : *"Natural Language Processing" AND (Accounting OR Accountability) Title* : *(text OR textual) AND (classifier OR classification) AND :(Multilabel OR "Multi-label" OR "Multi Label" OR Multiclass OR "Multi-class")*

Foram listados 13 resultados, dos quais 7 foram selecionados.

B.2.3 Inclusão e exclusão de estudos

Dos 37 estudos inicialmente selecionados nas sessões de buscas, apenas 2 foram excluídos da revisão. O primeiro estudo excluído foi um apêndice suplementar que estava duplicado, pois o trabalho ao qual ele se refere já continha o apêndice em sua versão integral. E de forma semelhante a metodologia adotada por Mirończuk e Protasiewicz (2018), que excluíram o repositório arXiv[5] em sua revisão sistemática, optou-se por excluir um trabalho publicado apenas como *pre-print* neste repositório, o que não atende ao critério de qualidade. Registra-se a ocorrência de inclusão de 1 estudo durante a fase de pesquisa exploratória, o trabalho de Zayas et al. (2017) foi incluído por ser um estudo de caso aderente aos critérios de inclusão 2, 3, 4 e 5 estabelecidos no protocolo da RSL, totalizando 36 trabalhos incluídos nesta RSL. Os Apêndices C.1 e C.2 detalham os estudos incluídos na revisão, assim como informações dos trabalhos excluídos constam no Seção C.3.

B.2.4 Extração de dados

A Figura 24 apresenta o formulário específico que foi elaborado para extração de dados dos estudos incluídos na RSL.

[5] https://arxiv.org/

Figura 24 – Formulário de extração de dados

Formulário de Extração de Dados – Ref. 000				
Título do trabalho:				
Abstract:				
Palavras-chave:				
Tipo de publicação:		**Fonte:**		
Data de publicação:		**Veículo publicação:**		
Tipo de trabalho:		**Idioma:**	Inglês	
Autores:	0	•		
Instituição:		**País:**		
Conceitos apresentados:	•			
Conceitos utilizados:				
Experimento?				
Resultados experimentais:				
Avaliação/Teste?				
Resultado avaliação:				
Técnicas apresentadas:	•			
Técnicas utilizadas:				
Linguagem de programação:				
Perspectiva humana: (É considerada? Como? Implica em algo?)				
Objeto pesquisa:				
Corpus de dados: (Nomes de bases conhecidas de dados)				
Origem dos dados: (Produzidos, Base corporativa, base publicada)				
Tipos de dados:				
Dados estruturados? (Sim, Não, Semi)				
Público-alvo dados: (Merc. Financeiro, Empresa, Gov., Púb., Pesq.)				
Aplicação: (Como pode ser utilizado o assunto do trabalho)				
QP1 Quais são os métodos e técnicas de ML que estão inseridos no contexto de processamento de registros textuais?				
QP2 Como é a utilização de ML para processamento de registros textuais?				
QP3 Qual a contribuição do ML, quando utilizada no processamento de textos, para área da visualização da informação?				
QP4 Quais são os conceitos e/ou tecnologias que se relacionam com a área de ML que possam se inter-relacionar com a área da visualização da informação?				
RESUMO:				
Trechos de destaque:				
Outras observações:				
Referências relevantes:				
Origem:	Pesquisa exploratória Google Scholar.			
Referência bibliográfica: *COLAR*				

Fonte: Produzida pelo autor.

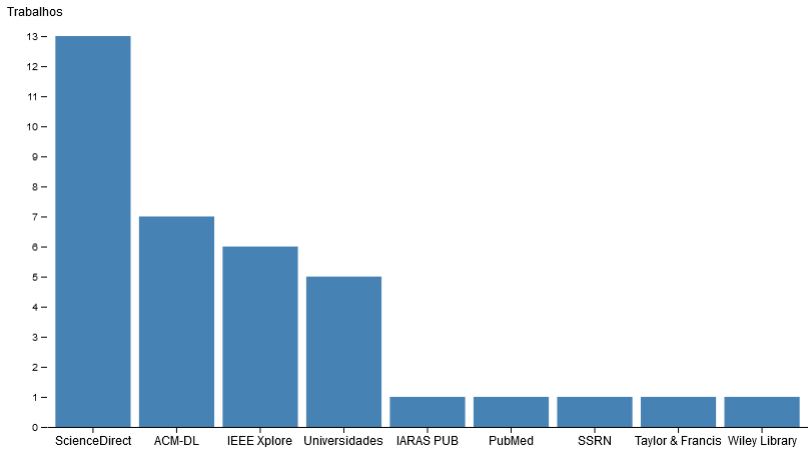
B.3 Relatório da Revisão Sistemática

Considerando os metadados dos trabalhos incluídos nesta RSL, em conjunto com dados objetivos e subjetivos extraídos durante a condução da revisão, esta seção discorre sobre aspectos gerais dos estudos, apresentando análise e discussão dos resultados.

B.3.1 Bases de publicações científicas

Dermeval, Coelho e Bittencourt (2019) destacam *ScienceDirect*[6], *ACM Digital Library*[7], *IEEE Xplore*[8], *SpringerLink*[9], *ISI Web of Science*[10], *Scopus*[11] e *Compendex (Engineering Village)*[12] como as principais bibliotecas digitais para área da Computação. Também destacam que dependendo da área de estudo, a interdisciplinariedade implica na utilização de bibliotecas digitais de outras áreas do conhecimento, citando a *PubMed Central*[13] como exemplo de base a ser utilizada para pesquisas em informática para educação médica. Neste sentido, optou-se por utilizar o Google Acadêmico, pois esta é uma base que oferece resultados próprios e indexados de outras fontes, o que auxilia na identificação de bases relevantes para o tema de pesquisa.

Figura 25 – Estudos publicados por Bases de Dados Científicas.



Fonte: Produzida pelo autor.

O gráfico da Figura 25 demonstra a distribuição de trabalhos por base de publicação, nota-se que majoritariamente os estudos incluídos nesta RSL foram publicados em bases recomendadas pelos autores (*ScienceDirect*, *IEEE Xplore* e *ACM Digital Library*), destacando-se em seguida os estudos obtidos em bibliotecas digitais universitárias indexadas pelo Google Acadêmico.

B.3.2 Veículos de publicação

Os estudos analisados nesta RSL foram publicados em periódicos, anais de conferências e bibliotecas digitais de universidades. A Tabela 18 apresenta detalhes relativos aos veículos de publicação e bases científicas.

[6] http://www.sciencedirect.com/
[7] http://dl.acm.org/
[8] http://ieeexplore.ieee.org/
[9] http://link.springer.com/
[10] http://apps.webofknowledge.com/
[11] http://www.scopus.com/
[12] http://www.engineeringvillage.com/
[13] https://www.ncbi.nlm.nih.gov/pubmed/

Tabela 18 – Bases e veículos das publicações.

Base	Veículo	Tipo	Publicações
ACM-DL	ACM International Conference on Information and Knowledge Management (CIKM'19)	C	1
	ACM Transactions on Management Information Systems	P	1
	Brazilian Symposium on Information Systems (SBSI'20)	C	1
	High Performance Computing and Cluster Technologies Conference (HPCCT'19)	C	1
	International ACM SIGIR Conference on Research and Development in Information Retrieval	C	1
	International Conference on Control Engineering and Artificial Intelligence (CCEAI'21)	C	1
	The Web Conference (WWW'20)	C	1
IARAS PUB	International Journal of Computers	P	1
IEEE Xplore	IEEE Access	P	2
	IEEE Computational intelligence magazine	P	1
	International Conference on Intelligent Computation Technology and Automation (ICICTA'18)	C	1
	International Conference on Promising Electronic Technologies (ICPET'18)	C	1
	International Joint Conference on Neural Networks (IJCNN'20)	C	1
PubMed Central®	Conference on Empirical Methods in Natural Language Processing (2018)	C	1
ScienceDirect	Applied Soft Computing	P	1
	Expert Systems With Applications	P	3
	Future Generation Computer Systems	P	1
	International Journal of Accounting Information Systems	P	1
	Journal of Biomedical Informatics	P	1
	Knowledge-Based Systems	P	2
	Neurocomputing	P	2
	Pattern Recognition	P	1
	The Journal of Finance and Data Science	P	1
SSRN	SSRN Electronic Journal	P	1
Taylor & Francis	International Journal of Computers and Applications	P	1
Universidades	Electronic Theses and Dissertations for Graduate School	D	1
	Open-Access-Publikationsserver der Humboldt-Universität	M	1
	Portal da Biblioteca Digital de Teses e Dissertações da USP	D	1
	Repositório UNIRIO	M	1
	Sydney Digital Theses (Open Access)	D	1
Wiley Online Library	Intelligent Systems In Accounting, Finance And Management	P	1

Tipos: P=Artigo de **P**eriódico(21) | D=Tese de **D**outorado(3) | M=Dissertação de **M**estrado(2) | C=Artigo de **C**onferência(10).

Fonte: Produzida pelo autor.

A análise dos veículos de publicação demonstra a interdisciplinariedade do tema Classificação de Texto, pois são registrados veículos de diversas áreas e temas distintos: Computação; Finanças; Contabilidade; Medicina; Linguagem; e, multidisciplinares.

B.3.3 Termos de indexação

Além das próprias palavras-chave atribuídas às publicações, alguns estudos apresentam termos específicos de indexação, que podem ser relacionados com a área de conhecimento ou atribuídos pela base de publicação. Huang e Li (2011) e Guo, Shi e Tu (2016) utilizam *JEL classification*[14], Young et al. (2018) utiliza *IET Inspec*[15] e trabalhos publicados na *ACM Digital Library* são indexados pelo *ACM CCS*[16]. No total foram encontradas 174 termos distintos utilizados para indexação dos estudos.

A nuvem de palavras exibida na Figura 26 foi construída utilizando-se os termos de indexação atribuídos aos artigos incluídos nesta RSL. Quanto maior o destaque da palavra na

[14] Journal of Economic Literature
[15] The Institution of Engineering and Technology
[16] ACM Computing Classification System

nuvem, maior a ocorrência de repetições de utilização desta em termos de indexação. Nota-se pela proeminência dos termos centrais da nuvem de palavras, que existe forte relação destas palavras com os termos utilizados nas sessões de buscas descritas nos tópicos B.2.1 e B.2.2.

Figura 26 – Palavras-chave relacionadas nos trabalhos selecionados na RSL.

Fonte: Produzida pelo autor.

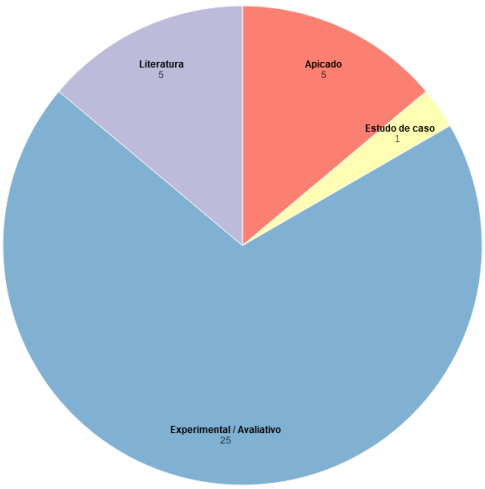
B.3.4 Tipos de trabalhos

A análise de características como objeto de pesquisa, tipo de abordagem, realização de experimentos e tipo de publicação, possibilitou categorizar os estudos em quatro categorias.

- Experimental/Avaliativo — Estudos que apresentaram realização de experimentos e análise comparativa entre técnicas, métodos ou modelos.
- Revisão da literatura — Mapeamento ou revisões sistemáticas.
- Aplicado — Trabalho teórico e experimental, categoria atribuída às teses e dissertações.
- Estudo de caso — Estudo analítico-descritivo da aplicação conjunta de técnicas e métodos.

A Figura 27 apresenta a proporção de trabalhos por categoria.

Figura 27 – Distribuição de estudos por tipo de trabalho.

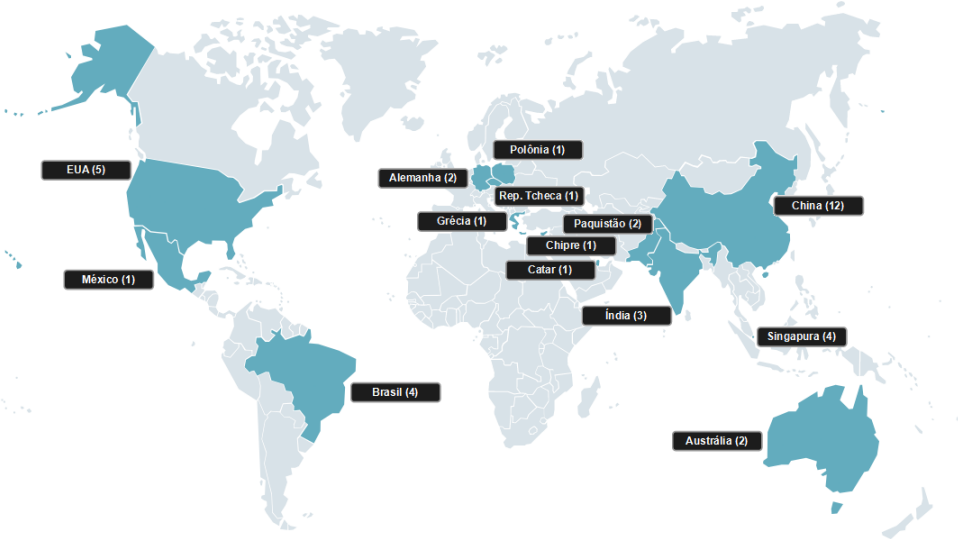


Fonte: Produzida pelo autor.

B.3.5 Distribuição geográfica

Ainda que a Classificação de Texto seja um tema antigo de pesquisa (FISHER; GARNSEY; HUGHES, 2016; KUMAR; RAVI, 2016), é evidente que ele permanece ativo e proeminente (MIROŃCZUK; PROTASIEWICZ, 2018). Afirmação esta que é reforçada pelo fato de que embora 36 estudos sejam uma amostra ínfima da literatura, ainda assim esta revisão sistemática apresenta trabalhos de 14 países que estão distribuídos em 5 continentes.

Figura 28 – Distribuição de estudos por nacionalidade de instituição de pesquisa.



Fonte: Produzida pelo autor.

A Figura 28 apresenta o país de origem das instituições de pesquisa nas quais autores estavam vinculados, conforme o registros de filiação informados nos trabalhos. Em parênteses, a quantidade de estudos oriundos do país. É notável a contribuição chinesa para o tema pesquisado, observa-se que 1/3 dos estudos incluídos nesta RSL são originários de autores do país. Merecem

destaque também os Estados Unidos da América, o Brasil e Singapura, que são os países que completam o pódio de maiores contribuições referenciadas nesta revisão sistemática.

B.3.6 Domínios de aplicação

De forma semelhante ao relatado por Mirończuk e Protasiewicz (2018), a Tabela 19 relaciona os domínios onde se aplica o objeto de pesquisa do estudo. A classificação de domínios de aplicação apresentada considera informações explicitadas nos trabalhos e domínios dos conjuntos de dados utilizados.

Tabela 19 – Domínios de aplicação dos estudos.

Domínio de aplicação	Estudos
Acadêmico/Científico	Mirończuk e Protasiewicz (2018) Schröder (2018) Young et al. (2018)
Contabilidade, Auditoria e Mercado Financeiro	Huang (2010) Huang e Li (2011) Fisher, Garnsey e Hughes (2016) Guo, Shi e Tu (2016) Kumar e Ravi (2016) Amani e Fadlalla (2017) Das, Mehta e Subramaniam (2017)
Informação e Conhecimento	Liu et al. (2017) Pink (2017) Giannopoulou e Mitrou (2018) Medeiros (2018) Tao, Cui e Wenjun (2018) Gong, Shi e Niu (2019) Raza et al. (2019) Bi et al. (2020) Mustafi, Mustafi e Sahoo (2020) Wang e Tan (2021) Xiao et al. (2021) Maltoudoglou et al. (2022)
Medicina	Rios e Kavuluru (2018) Ibrahim et al. (2021)
Mídias sociais	Li et al. (2016) Xiang e Zheng (2018) Cheng, Nazarian e Bogdan (2020)
Múltiplos domínios	Song (2009) Metz (2011) Huang et al. (2019) Zheng e Zheng (2019) Bittencourt, Silva e Almeida (2020) Santos e Merschmann (2020) Liu et al. (2021) Ma et al. (2022)
Energia/Eletricidade	Zayas et al. (2017)

Fonte: Produzida pelo autor.

B.3.7 Resultados Teóricos

Este seção apresenta uma visão geral de cada estudo classificado como revisão da literatura. É importante ressaltar que resultados detalhados destes trabalhos já foram apresentados, pois todo referencial teórico exposto nos Apêndices D e E foi baseado principalmente nestes artigos. Os estudos estão organizados aqui por ano de publicação e ordem alfabética de autor.

Fisher, Garnsey e Hughes (2016) realizaram estudo com objetivo de determinar o estado da extensa literatura de NLP aplicado em contabilidade, auditoria e finanças. Foram, analisados 262 estudos publicados entre 2010 e 2014, sendo 86 de técnicas manuais, 81 de mineração de texto básica (exemplos: NB, abordagem lexical, regras simples) e 95 de AI (exemplos: SVM, NN, SOM). Os autores constatam muitos desafios ainda abertos em relação a NLP aplicado ao domínio, principalmente na área de contabilidade, pois a maioria dos estudos e bases de dados são relacionados com o mercado financeiro.

Kumar e Ravi (2016) apresentam estudo com foco exclusivo no mercado financeiro, analisando 89 publicações realizadas entre 2000 e 2016. Os autores categorizaram os trabalhos de acordo com a aplicação identificada, criando quatro categorias: 1. Previsão de taxa de câmbio; 2. Previsão do mercado de ações; 3. Gerenciamento de relacionamento com clientes (CRM); e, 4. Segurança cibernética (exemplos: detecção de intrusão, fraude, *spam* e *phishing*). Identificando a classificação e a predição, como principais objetivos da Mineração de Texto no mercado financeiro, ao passo que problemas de segurança cibernética raramente utilizam esta abordagem. Observaram principalmente técnicas da abordagem *Machine Learning* no estudo, destacando o uso das técnicas SVM, NB, k-NN, DT, entre as quais SVM predomina o cenário de aplicação devido sua alta capacidade de predição.

Amani e Fadlalla (2017) analisam publicações de *Data Mining* relacionadas com a área de contabilidade, propondo um modelo de organização (*framework*) que correlaciona os três principais objetivos de DM (descrição, previsão e prescrição), com os dois principais aspectos de registros contábeis (retrospectiva e prospectiva). Os autores observam que de modo geral, as aplicações de DM em contabilidade tem perfil predominante relacionado a garantias e conformidade (*"assurance and compliance"*). Destacam que a previsão é o principal objetivo de aplicação, a classificação como tarefa mais frequente e, as redes neurais como principal técnica de implementação.

Young et al. (2018) apresentam um estudo de técnicas recentes de *Deep Learning*, aplicadas ao Processamento de Linguagem Natural. Os autores demonstram a evolução das técnicas relacionadas com a mineração de texto, partindo da representação do documento em *bag-of-words*, até representação distribuída para documentos de texto, denominada *word embeddingds*. A partir deste modelo de linguagem, discorrem sobre técnicas DL correlacionando-as com tarefas de NLP. Comentam que as representações distribuídas se tornaram os novos métodos de ponta para problemas de NLP, concluindo que DL é fundamental para construção e utilização de *word embeddingds*.

Mirończuk e Protasiewicz (2018) apresentam um profundo estudo com foco na Tarefa de Classificação de Texto, analisando 233 trabalhos publicados majoritariamente entre 2013 e 2018. Eles propõe um *framework* para organizar os trabalhos, que reflete sua visão do processo de

classificação de texto. Além disso, definem 8 Categorias Primárias, como objeto de estudo no tema Classificação de Texto: a) Sistemas de classificação e área de aplicação; b) Técnicas e métodos de rotulagem de amostras; c) Construção de características; d) Ponderação de características; e) Extração de características; f) Projeção de características; g) Método de classificação e métodos de aprendizagem; e, h) Avaliação da solução. Sendo [a,b,d,g,h] os temas com maior quantidade de trabalhos relacionados. Os autores observam que a representação do documento é vital para o processo de classificação, destacando *Word2Vec* e *GloVe* como principais ferramentas utilizada nesta etapa. Destacam ainda *Principal Component Analysis (PCA)*, *Singular Value Decomposition (SVD)* e *Linear Discriminant Analysis* como principais técnicas para redução de dimensionalidade do espaço de características. Algoritmos de classificadores não são citados nominalmente, descrevem que são variados e comentam que constam todos os paradigmas de aprendizagem, sendo a aprendizagem supervisionada a mais comum. Os autores concluem que a questão da classificação de texto ocorre em vários campos da atividade humana, que este é um assunto muito vivo na literatura atual, tendo como temas pouco explorados a classificação de múltiplas instâncias simultâneas, classificação de textos multilíngues e análise de fluxo de texto (*text stream analisys*).

B.3.8 Resultados Práticos

Esta seção discorre sobre os estudos que apresentaram resultados experimentais e avaliações de métodos e técnicas de aplicação. Constam nesta seção 15 publicações de periódicos, 10 artigos de conferências, 3 teses, 2 dissertações e 1 estudo de caso. A discussão sobre os trabalhos está organizada por semelhança entre os estudos e pelo tipo de problema de classificação abordado por eles.

B.3.8.1 Balanceamento Amostral

Chawla et al. (2002) propõem a técnica SMOTE (do inglês, *Synthetic Minority Oversampling Technique*), para tratamento de classes desbalanceadas no conjunto de treinamento. O algoritmo SMOTE cria amostras sintéticas a partir das classes minoritárias, para que estas classes tenham exemplos de treinamento em quantidade equivalente à classe com maior número de amostras. Raza et al. (2019) aplicam o SMOTE em seu estudo na fase de pré-processamento e utilizam as amostras geradas para treinar métodos de *Machine Learning* para atuar com problemas de classificação monorrótulo. Já Xiang e Zheng (2018) realizam estudo avaliativo sobre a utilização do SMOTE com técnicas de *Deep Learning*, nominalmente *Word Embedding* e Rede Neural Convolucional, aplicadas ao enfrentamento de problemas de classificação multirrótulo. Xiang e Zheng concluem que o paradigma de métodos proposto apresenta expressivo aumento na acurácia medida na tarefa de classificação multirrótulo, quando comparado com métodos de ML e/ou com representação vetorial de documentos.

B.3.8.2 Problemas de Classificação Binária

O artigo de Guo, Shi e Tu (2016) é uma pesquisa bibliográfica e aplicada, com foco na relação de ML com estudos financeiros. O estudo se inicia discutindo duas técnicas de abordagem lexical, onde os pesquisares comentam sobre o conceito de Legibilidade e suas métricas, realizando

em seguida experimentos com os dicionários léxicos *Harvard General Inquirer* e *Loughran and Mcdonald*. Os autores analisaram quatro técnicas de processamento de registros textuais, com intuito de classificar notícias relacionadas à empresas do mercado financeiro norte-americano. Três técnicas foram experimentadas para classificação binária: i) *Naive Bayes (NB)*; *Support Vector Machine (SVM)*; e, *Back Propagation Neural Network (BPN)*. Das técnicas apresentadas, as redes neurais demonstraram melhor desempenho na classificação.

A classificação binária também está presente em dois módulos do sistema de rastreamento de boatos e *fake news* desenvolvido por Cheng, Nazarian e Bogdan (2020). O módulo detector de boatos utiliza esse modelo de classificação para determinar se a amostra de texto é boato ou não é boato. E o módulo rastreador de boatos identifica se o boato já foi ou não relatado.

B.3.8.3 Problemas de Classificação Monorrótulo de Texto

Problemas de classificação multiclasse-monorrótulo são os mais frequentes na literatura relacionada com abordagens de *Machine Learning*. Além das técnicas relacionadas com a classificação binária, o estudo de Guo, Shi e Tu (2016) apresenta o algoritmo de classificação semântica *Latent Dirichlet Allocation (LDA)*, que mostrou-se muito interessante para a descoberta de novas informações, como a relação entre as classes.

Huang (2010) e Das, Mehta e Subramaniam (2017) desenvolveram algoritmos próprios utilizando a abordagem de transformação de problema *Binary Relevance (BR)*, Guo, Shi e Tu (2016) também utiliza BR como abordagem, mas implementa com algorítimo de classificação semântica *Latent Dirichlet Allocation (LDA)*. Implementações com métodos básicos de ML estão presentes em Song (2009) e Raza et al. (2019), com abordagem de transformação *One-vs-All*. Na literatura relacionada com abordagens de *Deep Learning*, destacam-se os estudos de Rios e Kavuluru (2018) e Zheng e Zheng (2019), por atuarem com problemas monorrótulo utilizando redes neurais complexas com *Attention Mechanism* e *Stacking* respectivamente.

Ainda na literatura relacionada com abordagens de *Machine Learning*, mas lidando com problemas de classificação multirrótulo, Li et al. (2016) apresentam modelo de classificação baseado no conceito de máxima entropia, para relacionar palavras de texto curto com rol de sentimentos elegíveis. Santos e Merschmann (2020) introduzem o conceito de *metalearning* e propõe a abordagem de seleção dinâmica e automática de método de classificação, de acordo com a similaridade da amostra em relação ao histórico de melhores classificadores para aquele tipo de amostra.

Bittencourt, Silva e Almeida (2020) discorrem sobre o modelo autoral denominado ML-MDLText. O modelo proposto baseia-se no princípio do Comprimento Mínimo de Descrição (MDL, do inglês, *Minimum Description Length*) e o modelo destaca-se por duas características distintas dos demais métodos de ML vistos nesta RSL: i) o problema de classificação multirrótulo é abordado sem a utilização das técnicas de transformação de problema ou adaptação de algoritmo; e, ii) o modelo de classificação apresenta um método online e interativo de aprendizado. O método apresentado no estudo foi inspirado em trabalhos anteriores dos autores, a principal diferença entre é que esta versão do ML-MDLText foi concebida para enfrentar o problema de classificação multirrótulo. A principal ideia do modelo proposto é utilizar conceitualmente

características das técnicas *Binary Relevance (BR)* e *Label Powerset (LP)* para solucionar o problema de classificação multirrótulo, sendo que BR inspira a utilização de informações de ocorrências de termos para cada classe individualmente e LP inspira que informações relacionadas às ocorrências de termos sejam agregadas a cada conjunto de rótulos elegíveis. Desta forma o modelo de classificação se beneficia pelo uso das informações sobre as classes e da relação de proximidade semântica entre os rótulos.

B.3.8.4 Problemas de Classificação Multirrótulo de Texto

Problemas de classificação multiclasse-multirrótulo de texto são abordados no trabalho de Xiao et al. (2021). Os autores destacam o recente crescimento desta área de pesquisa e os modelos atualmente considerandos como estado da arte para tal tarefa de classificação. Considerando notável a contribuição dos mecanismos de atenção para a tarefa de classificação multirrótulo, os autores apresentam trabalho com objetivo de incrementar estes mecanismos para considerar mais relações na modelagem de documentos realizada pelo método generativo *sequence-to-sequence (Seq2Seq)*, com objetivo de aumentar a quantidade de rótulos aplicados aos textos analisados e diminuir a ocorrência de atribuição de rótulos errados. O modelo proposto demonstrou-se superior a todos os dez modelos referenciados como estado da arte na classificação multirrótulo de texto, sendo que ficou em primeiro lugar nas métricas *Recall* e *Hamming Loss*. Contudo, o modelo proposto obteve o segundo lugar considerando a métrica *Precision*.

Maltoudoglou et al. (2022) partem de um trabalho anterior onde apresentaram o modelo denominado ICP (do inglês, *Inductive Conformal Prediction*), os autores demonstram no trabalho atual a nova abordagem chamada de LP-ICP. Apresentam neste trabalho de forma detalhada, tanto conceitual quanto tecnicamente, sobre como abordar limitações computacionais da técnica *Label Powerset (LP)*, propondo uma implementação que reduz o custo computacional para utilização da técnica LP em problemas de classificação multirrótulo de texto. Os autores estabelecem matematicamente a validade da abordagem proposta e comprovam a eficiência do método com resultados experimentais, fornecendo inclusive resultados experimentais para problemas onde anteriormente era computacionalmente desafiador.

Tao, Cui e Wenjun (2018) O estudo propõe uma nova abordagem para problemas de classificação multirrótulo, utilizando uma fusão de rótulos criada por meio de rede neural convolucional, com objetivo de aumentar a quantidade de rótulos possíveis a serem atribuídos ao texto, considerando a relação semântica entre os rótulos em vez de considerar relações estatísticas. A relação de mapeamento entre o texto e o vetor do rótulo é construída por meio da CNN e a saída desta rede é usada como o vetor de rótulo do texto para recuperar os rótulos vizinhos mais próximo no conjunto de rótulos. O método proposto enriquece o conjunto de rótulos de texto atribuídos às amostras de texto, incrementando-os com a utilização do conceito do vizinho mais próximo da semântica textual. Resultados do estudo demonstram que o modelo proposto é superior a técnicas como LP, BR e ML-KNN para a extensão de rótulos semânticos de texto.

Os trabalhos de Huang e Li (2011) e Schröder (2018) são os únicos estudos destacando multirrótulo que utilizam métodos de *Machine Learning*. Huang e Li (2011) utilizam abordagem *Binary Relevance (BR)* e propõe o método *Multilabel Categorical K-Nearest Neighbor (ML-*

CKNN). Schröder (2018) desenvolve métodos baseados em *Latent Dirichlet Allocation (LDA)* que utilizam abordagem de transformação *One-vs-All*.

B.3.8.5 Problemas de Classificação Multirrótulo Extrema de Texto

Jingzhou Liu et al. (2017) apresentam o conceito de classificação multirrótulo extrema de texto (XMTC, do inglês, *Extreme Multi-label Text Classification*), onde o tamanho do conjunto de rótulos é muito grande, contendo centenas de milhares ou milhões de rótulos disponíveis. No estudo apresentado, os autores propõe uma nova abordagem para o problema XMTC, utilizando de aprendizado profundo de máquina para a tarefa de classificação multirrótulo extrema, com objetivo de avaliar o desempenho e escalabilidade da solução proposta baseada em *Deep Learning*.

Huiting Liu et al. (2021) também apresentam estudo sobre a viabilidade computacional para solução de problemas de classificação de texto multirrótulo, em corpus com grande quantidade de documentos e rótulos. E Ibrahim et al. (2021) propõem modelo para classificação multirrótulo híbrido, que funciona bem tanto para classificação extrema quanto para classificação tradicional de textos.

Ibrahim et al. (2021) apresentam estudo da utilização conjunta de redes neurais convolucionais e recorrentes, aplicadas à tarefa de classificação multirrótulo de texto no domínio da biomedicina. Demonstrando bons resultados com seu modelo GHS-Net (do inglês, *Generic Hybridized Shallow Neural Network*), os autores afirmam que o modelo proposto é extremamente eficiente para este domínio do conhecimento. O conceito de generalização exposto no título do artigo é devido ao fato do mesmo modelo poder ser utilizado tanto em corpus estruturados da literatura biomédica, quanto em corpus de notas clínicas feitas em linguagem natural. Ou seja, o modelo proposto atente a tarefa de classificação multirrótulo extrema no corpus da literatura biomédica e também atende a tarefa de classificação multirrótulo "normal".

B.3.8.6 Problemas de Classificação Hierárquica de Texto

Medeiros (2018) e Metz (2011) trabalham métodos de *Machine Learning* e implementam algoritmos próprios utilizando a abordagem *Binary Relevance (BR)* para solução de problemas de classificação hierárquica, que também podem ser considerados monorrótulo, pois só permitem classificação de uma instância em um ramo de generalização/especialização.

B.3.8.7 Problemas de Classificação Hierárquica e Multirrótulo de Texto

Huang et al. (2019) apresentam a classificação hierárquica e multirrótulo de texto (HMTC, do inglês, *Hierarchical Multi-label Text Classification*), como uma importante e desafiadora tarefa para lidar com aplicações e documentos do mundo real. Os autores propõem um modelo de classificação hierárquica de documentos, levando em consideração as categorias mais relevantes, nível por nível da estrutura hierárquica da árvore de categorias disponível. O modelo proposto é estruturalmente composto por três partes, cada qual com sua função e objetivo próprio: Camada de Representação de Documentação (DRL); Camada Recorrente Baseada em Atenção Hierárquica (HARL); e, Camada de Predição Híbrida (HPL). Primeiro utiliza-se a DRL para obter a representação unificada de cada texto de documento e a estrutura de categoria hierárquica,

aplicando-se a técnica de *Word Embedding*. Em seguida, a HARL modela as dependências entre os diferentes níveis de categorias hierárquicas, capturando associações entre os textos e cada categoria da estrutura hierárquica de cima para baixo. Por fim, a HPL é aplicada para predição das categorias hierárquicas relacionadas com o documento.

Bi et al. (2020) demonstram o *framework* chamado Modelo de Codificação Global-Local (GLEN, do inglês, *Global-Locally Encoding*), que é composto por três módulos: Módulo de Codificação Global, Módulo de Codificação Local e Módulo de Pontuação. O GLEN realiza a extração de informações tanto de escopo global do texto, quanto informações específicas (locais) para cada categoria disponível no espaço de rótulos. As informações extraídas são utilizadas na tarefa de classificação de texto, através do ranqueamento das pontuações ponderadas por escopo. A implementação do modelo proposto utiliza uma combinação da técnica de *Binary Relevance* em conjunto com redes neurais, onde segundo os autores "alguns classificadores relativamente independentes são aprendidos por BR, e a capacidade de ajuste da rede neural também é utilizada para extrair algumas informações globais e locais". Ou seja, o problema de classificação multirrótulo de texto é decomposto em classificação binária para cada rótulo e utilizam-se redes neurais recorrentes e convolucionais para extrair informações do texto. O modelo proposto demonstrou-se eficiente em relação aos métodos consagrados como estado da arte na tarefa de classificação multirrótulo de texto.

Ma et al. (2022) apresentam o HE-HMTC (do inglês, Hybrid Embedding based text representation for HMTC), um modelo híbrido de representação de documentos de texto, que pode ser aplicado à tarefa de classificação de texto hierárquica e multirrótulo. O método proposto consiste na junção de um modelo de representação estrutural com um modelo de representação de palavras. A representação estrutural implementada é baseada em *Structural Deep Network Embedding (SDNE)*, uma rede de incorporação estrutural que organiza em grafos acíclicos a estrutura hierárquica das categorias. Ao passo que a implementação da representação de palavras é efetuada com *Word Embedding*, que incorpora em uma representação vetorial cada categoria representada pelos nós do grafo. Os autores realizaram experimentos para validação do modelo proposto utilizando cinco conjuntos de dados, comparando o HE-HMTC com abordagens consideradas o estado da arte para tarefa de classificação hierárquica e multirrótulo de texto. Nota-se que o modelo proposto apresenta resultados competitivos tanto na comparação com classificadores planos adaptados, quanto na comparação com outros classificadores hierárquicos. Sendo que o HE-HMTC muitas vezes obteve a melhor precisão geral nos conjuntos de dados com níveis hierárquicos mais profundos.

B.3.8.8 Modelagem de Documentos e Mecanismos de Atenção

Dentre os resultados de aplicações práticas já descritos, como os trabalhos de Bi et al. (2020), Ma et al. (2022), Maltoudoglou et al. (2022), Xiao et al. (2021), nota-se grande preocupação dos pesquisadores no estudo de novas técnicas para representação de documentos de texto. Os autores aplicam de forma integrada métodos já consagrados e técnicas customizadas para criarem seus próprios modelos de representação de documentos.

Da mesma forma que pesquisadores buscam aperfeiçoar os modelos de representação de

documentos, também têm buscado novas formas de integrar mecanismos de atenção aos seus modelos de classificação. Huang et al. (2019) aplica mecanismos de atenção para identificar rótulos mais relevantes em cada nível da estrutura hierárquica de classes elegíveis para classificação. Maltoudoglou et al. (2022) buscam incrementar os mecanismos de atenção, para considerar mais relações na modelagem de documentos *sequence-to-sequence*.

Gong, Shi e Niu (2019) apresentam uma nova forma de implementação de Redes de Auto-atenção (SAN, do inglês, *Self-Attention Network*), aplicada à tarefa de classificação de textos em linguagem natural. Utilizando-se da divisão do texto em pequenas sentenças e realizando a estruturação do documento de forma hierárquica, o método proposto consegue capturar efetivamente dependências em longas sequências de documentos e superar o grande requisito de memória dos métodos de auto-atenção existentes. Além disso, o estudo utiliza o espaço de rótulos como parte integrada do mecanismo de auto-atenção, utilizando-os na construção de representações dos documentos. Os experimentos realizados demonstram a eficácia na tarefa de classificação de documentos, considerando a precisão e os requisitos de memória.

Cheng, Nazarian e Bogdan (2020) abordam a aplicação de NLP para detecção e rastreamento de boatos e *fake news* em mídias sociais. Utilizando a modelagem de documentos generativa, apresentam sistema com quatro módulos, que contemplam diferentes etapas do processo de detecção de boatos. Sendo que cada módulo dedica-se ao enfrentamento de problemas de classificação relacionados com sua respectiva etapa: i) detector de boatos; (ii) rastreador de boatos; (iii) classificador de instância; e, (iv) classificador de veracidade. O sistema apresentado mostrou-se eficiente e superou técnicas consagradas na aplicação de detecção de boatos.

Liu et al. (2021) propõe o modelo de classificação multirrótulo denominado LELC (*Label Embedding and Label Correlation*), para resolver o problema de classificação de texto multirrótulo com muitas classes. O modelo aplica o conceito *Label Space Dimension Reduction* (LSDR) para redução da dimensionalidade do espaço de rótulos. O método apresentado utiliza um mecanismo de atenção multicamadas na codificação/incorporação de rótulos (*Label Embedding*) e leva em conta a correlação destes rótulos para realizar a decodificação dos rótulos atribuídos. Resultados experimentais em 11 conjuntos de dados do mundo real demonstram a eficácia do modelo proposto.

Wang e Tan (2021) propõe a representação de documentos de texto baseada em rótulos (das classes elegíveis para classificação), utilizando uma rede neural convolucional baseada em rótulos (*LBCNN - Label Based CNN*). O modelo apresentado realiza a atribuição de peso às palavras que compõe o texto, baseando-se no conjunto de rótulos disponíveis e utilizando recursos semânticos do texto. O método foi testado em grandes conjuntos de dados consagrados na literatura. Resultados experimentais demonstram que o LBCNN atinge desempenho comparável ou superior na classificação de texto, quando comparado com outros modelos de referência. Wang e Tan destacam que o modelo proposto tem baixo custo computacional, afirmando ainda que os estudos demonstram que a informação dos rótulos das classes é muito importante, podendo até substituir os mecanismos de atenção na ponderação de relevância de palavras. Apesar de todos os benefícios citados, os autores apontam duas deficiências do modelo: i) a alta dependência de vetores de palavras pré-treinados; e, ii) o fato de que a representação de texto baseada em rótulos

afeta diretamente a tarefa de classificação, mas que este modo de representação de documentos ainda é pouco explorado.

B.3.8.9 *Clustering*

Mustafi, Mustafi e Sahoo (2020) sugerem um novo método clusterização, que afirmam ser idealmente ajustado para formar clusters de documentos de texto. Método baseado em *Nearest Neighbour Separation (NNS)*, aplicando método eurístico para descoberta de vizinhos mais próximos, com a finalidade de tunar a função de adequação do Algoritmo Genético (GA) de agrumamento. Tendo a abordagem tradicional do algoritmo *K-Means* como base, os autores afirmam que o método proposto tem desempenho superior e custo computacional pouca coisa mais alta, compensado pelo fato de que para demandas de agrupamento altamente especializados, o método oferece uma separação mais precisa quando comparado ao algoritmo base.

Giannopoulou e Mitrou (2018) implementam diversos modelos de *Self-Organizing Maps (SOM)* aplicados a solução de problemas multirrótulo, com a utilização da abordagem de transformação de problemas *Label Powerset (LP)*. O método desenvolvido realiza uma abordagem em duas etapas na construção do vetor que alimenta o SOM. Na primeira etapa gera-se um vetor onde cada amostra contém um conjunto de atributos e um conjunto de rótulos criados utilizando os conjuntos de treinamento do vetor. Enquanto na segunda etapa quebra-se o vetor de rótulos, gerando cópias da mesma amostra para cada rótulo atribuído à ela. Uma característica interessante do trabalho é o desenvolvimento de um método mais genérico possível, pois sua extração de características automatizada se adequaria a qualquer tipo de *dataset*.

B.3.8.10 *Slot Filling*

Pink (2017) apresenta o estudo menos relacionado com o tema de pesquisa desta RSL. O autor discorre em sua tese sobre o tópico *Slot Filling (SF)* e sua relação com *Relation Extraction (RE)*. Apesar de abordar o tema Extração de Relação de forma ampla e com muitos exemplos, inserindo o SF no contexto, a relação com técnicas de *Machine Learnig* ligadas ao campo de estudo é apresentada de forma superficial, comentando mais sobre as abordagens de aprendizagem. Mas, relacionando SF como uma das aplicações possíveis para ML.

B.3.8.11 Estudo de Caso

Por fim, Zayas et al. (2017) apresentam um estudo de caso que relaciona todos conceitos abordados nesta RSL. Os autores conceitualizam *Analytics* como um processo de análise sistemática de dados, que utiliza várias técnicas para obter *insights* de um conjunto de dados. Comentam que as técnicas de *analytics* são baseadas na combinação de regras de negócios, análise estatística, algoritmos, ML, DM, NLP, Análise de Texto, AI, Visualização da Informação e outros. O artigo apresenta o estudo da implementação de uma plataforma analítica de dados, que se enquadra na perspectiva *Big Data*. Os autores comentam que esta plataforma analítica poderia ser implementada através de soluções corporativas adquiridas ou que podem ser desenvolvidas por conta da própria instituição. Neste trabalho os autores realizaram o desenvolvimento da plataforma analítica utilizando o *Apache Hadoop Framework*, concluindo que o desenvolvimento

é custoso em relação a recursos humanos, financeiros e de tempo, mas que as informações disponibilizadas através deste tipo de plataforma vão além do conceito de *Business Intelligence (BI)*, culminando com a transformação da empresa e seus processos. Não houve comparação de vantagens ou desvantagens em relação a compra ou desenvolvimento próprio da solução, apenas foi comentado que mesmo as soluções compradas demandam desenvolvimento para adequação e transformação dos sistemas legados.

B.3.9 Sumarização de Resultados

A análise dos resultados teóricos e práticos registrados, em conjunto com as Questões de Pesquisa (QP) que nortearam esta revisão sistemática, a saber:

1. Quais são os métodos e técnicas de Aprendizado de Máquina que estão inseridos no contexto de processamento de registros textuais?
2. Como é a utilização do Aprendizado de Máquina para processamento de registros textuais?
3. Qual a contribuição do Aprendizado de Máquina, quando utilizado no processamento de textos, para área da Visualização da Informação?
4. Quais são os conceitos e/ou tecnologias que se relacionam com as áreas de Inteligência Artificial e Aprendizado de Máquina, que possam se inter-relacionar com a área da Visualização da Informação?

É possível sumarizar os resultados obtidos, relacionando os trabalhos analisados com as questões de pesquisa estabelecidas. A Tabela 20 demonstra esta relação.

Tabela 20 – Relação de trabalhos com as Questões de Pesquisa (QP).

	QP1	QP2	QP3	QP4
Amani e Fadlalla (2017)	ML	*Data Mining*	*Clustering*	*Visual analytics*
Bi et al. (2020)	DL	Classificação hierárquica e multirrótulo + NLP	—	—
Bittencourt, Silva e Almeida (2020)	ML	Classificação multirrótulo	—	—
Cheng, Nazarian e Bogdan (2020)	DL	NLP	—	—
Das, Mehta e Subramaniam (2017)	ML	Classificação	—	—
Fisher, Garnsey e Hughes (2016)	ML + DL + AI	NLP	—	*Visual analytics*
Giannopoulou e Mitrou (2018)	AI (SOM)	*Clustering* + classificação	*Clustering*	*Visual analytics*
Gong, Shi e Niu (2019)	DL	NLP + Classificação	—	—
Guo, Shi e Tu (2016)	ML + Semântica:LDA	Classificação binária + NLP	—	*Visual analytics*
Huang (2010)	ML	NLP	—	—
Huang e Li (2011)	ML	NLP	—	—
Huang et al. (2019)	DL	Classificação hierárquica	—	—
Ibrahim et al. (2021)	ML + DL	Classificação multirrótulo tradicional/extrema + NLP	—	—
Kumar e Ravi (2016)	ML	Classificação	Estruturação de texto	CRM + BI
Li et al. (2016)	ML	Classificação de texto curto	—	—
Liu et al. (2017)	DL	Classificação multirrótulo extrema	—	—
Liu et al. (2021)	ML + DL	Classificação multirrótulo + NLP	—	—
Ma et al. (2022)	DL	Classificação hierárquica e multirrótulo + NLP	—	—
Maltoudoglou et al. (2022)	DL	Classificação multirrótulo + NLP	—	—
Medeiros (2018)	ML	Classificação hierárquica	—	—
Metz (2011)	ML	Classificação hierárquica	—	—
Mirończuk e Protasiewicz (2018)	ML	Várias aplicações	—	—
Mustafi, Mustafi e Sahoo (2020)	AI (GA)	*Clustering*	*Clustering*	*Visual analytics*
Pink (2017)	—	Slot Filling	—	—
Raza et al. (2019)	ML	Classificação	—	*Visual analytics*
Rios e Kavuluru (2018)	DL	Classificação	—	*Visual analytics*
Santos e Merschmann (2020)	ML + *Metalearning* + Semântica:*Keyword detection*	Classificação multirrótulo	—	—
Schröder (2018)	Semântica:LDA	NLP	Estruturação de texto	*Visual analytics*
Song (2009)	ML	Classificação + recomendação	*Clustering*	*Visual analytics*
Tao, Cui e Wenjun (2018)	DL	Classificação multirrótulo + NLP	—	—
Wang e Tan (2021)	DL	NLP + Classificação	—	—
Xiang e Zheng (2018)	DL	NLP + Classificação	—	—
Xiao et al. (2021)	DL	NLP + Classificação	—	—
Young et al. (2018)	DL	Classificação	—	—
Zayas et al. (2017)	ML + DL + NLP + AI	*Data Mining*	*BI + Dashboards*	*Analytics + BI*
Zheng e Zheng (2019)	DL	Classificação refinada de texto	—	*Visual analytics*

Fonte: Produzida pelo autor.

B.3.10 Considerações Finais do Relatório da RSL

Analisou-se nesta seção os resultados obtidos com a Revisão Sistemática da Literatura. Os resultados literários e práticos desta RSL, revelam que não há um método ou técnica de AI que seja indiscutivelmente superior, mas existe consenso que abordagens mais modernas e sofisticadas frequentemente apresentam melhor desempenho na tarefa de classificação. Desta forma, é possível inferir que de modo geral os métodos de *Deep Learning* são superiores aos modelos de *Machine Learning*. Considerando apenas métodos de ML, técnicas baseadas em *Logistic Regression (LR)*, *Passive Aggressive (PA)* e *Support Vector Machine (SVM)* costumam apresentar os melhores desempenhos. Além dos conceitos, técnicas e métodos descritos nos resultados teóricos e práticos, destacam-se a utilização da representação de documentos em formato vetorial, normalmente relacionadas com abordagem ML e criados com a técnica *TF–IDF*. *Word Embedding* atualmente é o método de referência para modelagem de documentos em DL. A análise de relações estabelecidas na Tabela 20 evidencia a forte relação entre *Natural Language Processing (NLP)* e *Deep Learning*, mas destacam-se abordagens semânticas em ML utilizando-se dicionários, palavras-chave e *Latent Dirichlet Allocation (LDA)*.

B.4 Análise, Discussão e Conclusão da RSL

Esta seção dedica-se à análise dos resultados obtidos nesta revisão sistemática, apresenta observações realizadas durante o estudo e encerra a RSL com suas contribuições para o projeto de pesquisa.

B.4.1 Observações Realizadas Durante o Estudo

Alguns detalhes como a classificação manual de texto, diferenças na forma de escrita de conceitos abordados, diferentes nomenclaturas para o mesmo conceito e a ausência de alguma abordagem específica de conceitos, foram assuntos que se evidenciaram e merecem ser destacados.

B.4.1.1 Classificação Manual de Texto

O estudo de Fisher, Garnsey e Hughes (2016), foi o primeiro contato da literatura que abordou a classificação manual de texto, destacando-se por apresentar quase 1/3 dos trabalhos analisados utilizando o modo operante manual.

Medeiros (2018) apresenta uma abordagem interessante para realização de classificação de textos de forma manual. A utilização de plataformas de *crowdsourcing*, que permite a contratação de centenas ou milhares de pessoas para execução da tarefa. Este método foi utilizado pelo autor para validar seu modelo de classificação automática proposto.

B.4.1.2 Alternativas Para Análise de Texto

Trabalhos mais antigos remetem aos primórdios da análise de texto, utilizando conceitos e técnicas que antecedem a aplicação de métodos de *Machine Learning*. Ingram e Frazier (1980 apud GUO; SHI; TU, 2016) introduz a análise de texto utilizando da contagem de frequência de palavras-chave.

Huang e Li (2011) agrupa resumidamente técnicas de análise de texto, dividido-as em quatro categorias, considerando seus métodos para quantificação de informações textuais: (1) análise de conteúdo usando software empacotado (software pronto); (2) contagem de palavras-chave; (3) classificação de texto; e, (4) Processamento de Linguagem Natural (NLP).

Fisher, Garnsey e Hughes (2016) comentam sobre trabalho cujo domínio é o mercado financeiro, onde realiza-se aplicação conjunta de NLP com a técnica *Keyword in Context (KWIC)*, para extrair e categorizar informações de relatórios fiscais. Neste mesmo domínio de aplicação, Das, Mehta e Subramaniam (2017) afirmam que dicionários de palavras-chave escolhidas manualmente por especialistas do domínio, são frequentemente utilizadas para extração de informações de textos do mercado financeiro.

Mirończuk e Protasiewicz (2018) citam a classificação baseada em palavras-chave, como um dos temas de pesquisa na área de classificação de texto, principalmente para pesquisas com foco na fase de pré-processamento (e.g., *feature construction*).

Santos e Merschmann (2020) também comentam sobre a fase de pré-processamento (e. g., *attribute Extraction*), discorrendo sobre a análise semântica utilizando-se dicionários e aplicando técnica de detecção de palavras-chave.

B.4.1.3 Mesmo Conceito, Diferentes Escritas

A escolha dos termos de busca se mostrou eficiente ao considerar diferentes formas de escrita dos termos de interesse, pois foram encontrados exemplos de diferentes escritas realizadas pelo mesmo autor:

- Huang (2010, grifo nosso) intitula seu trabalho como *"Exploring the information contents of risk factors in SEC form 10-K: A **multi-label** text classification application"*. Em seguida, Huang e Li (2011, grifo nosso) apresentam o trabalho denominado *"A **multilabel** text classification algorithm for labeling risk factors in SEC form 10-K"*.
- Schröder (2018, grifo nosso) intitula sua dissertação como *"Hierarchical **Multiclass** Topic Modelling with Prior Knowledge"*. E ao discorrer sobre o problema comenta: *"Our setting is one of **multi-class** classification in which every JEL leaf label is considered its own class"*.

B.4.1.4 Diferenças Conceituais

A nomenclatura de problemas de classificação apresenta bastante variação na literatura. Aparentemente as diferenças de língua materna e língua escrita dos trabalhos contribui tanto para variações como para confusões.

Os termos em inglês *multi-class* e *multi-label*, foram traduzidos neste trabalho respectivamente para multiclasse e multirrótulo. Pois é desta forma que foram aplicados em grande parte dos estudos analisados: Song (2009), Huang (2010), Huang e Li (2011), Li et al. (2016), Liu et al. (2017), Giannopoulou e Mitrou (2018), Tao, Cui e Wenjun (2018), Xiang e Zheng (2018), Huang et al. (2019), Mirończuk e Protasiewicz (2018), Raza et al. (2019), Zheng e Zheng (2019), Bi

et al. (2020), Bittencourt, Silva e Almeida (2020), Santos e Merschmann (2020), Ibrahim et al. (2021), Xiao et al. (2021), Ma et al. (2022) e Maltoudoglou et al. (2022).

Destacam-se os trabalhos de Rios e Kavuluru (2018) e Medeiros (2018) que apresentam divergências conceituais:

Medeiros (2018) utiliza o termo *multi-class* para denotar o problema de classificação multirrótulo.

> *Two different types of text categorization task can be identified depending on the number of categories that could be assigned to each document. The first type, in which precisely one category is assigned to each $d_j \in D$, is named as the single-class (or non-overlapping categories) text categorization task. The second type, in which* ***any number of categories from zero to*** *$|C|$* ***may be assigned to each*** *$d_j \in D$****, is called the multi-class (or overlapping categories)*** *task (MEDEIROS, 2018, grifo nosso).*

Rios e Kavuluru (2018) apresenta logo no *abstract* o termo *multi-label*, para se referir ao conjunto de várias classes, que denota um problema de classificação multiclasse.

> ***Large multi-label datasets contain labels that occur thousands of times*** *(frequent group), those that occur only a few times (few-shot group), and labels that never appear in the training dataset (zero-shot group). Multi-label few- and zero-shot label prediction is mostly unexplored on datasets with large label spaces, especially for text classification (RIOS; KAVULURU, 2018, grifo nosso).*

B.4.1.5 Ausência de Métodos *Fuzzy*

Kumar e Ravi (2016) destacam em seu estudo a ausência de métodos de inteligência artificial relacionados com abordagem *Fuzzy*. Nesta RSL também não foram encontrados estudos com este modelo. O único trabalho de revisão da literatura que registrou presença deste método, foi o estudo de Fisher, Garnsey e Hughes (2016), onde apenas 2 artigos foram listados com esta abordagem.

B.4.1.6 Premissas Sobre Problemas Multiclasse e Multirrótulo

A premissa "todo problema multirrótulo é um problema multiclasse", aparentemente é verdadeira. Porque não faz sentido a possibilidade de se atribuir várias categorias, quando o conjunto de classes possíveis é de ordem binária. Um contraponto quanto a veracidade da premissa pode ser observado no trabalho de Medeiros (2018), que define a classificação multiclasse utilizando o intervalo de classes de 0 até n. Neste caso, faz sentido um problema onde o conjunto de classes é de ordem binária ser chamado de problema multiclasse, pois existe a possibilidade de não ser atribuída nenhuma classe.

Contudo, os estudos desta RSL revelaram que não é válido considerar a premissa "todo problema multiclasse é um problema multirrótulo", pois existem problemas multiclasse com categorias mutuamente excludentes.

B.4.2 Discussão

A discussão final deste estudo contempla as respostas das questões de pesquisa estabelecidas na Revisão Sistemática da Literatura.

1. Quais são os métodos e técnicas de Aprendizado de Máquina que estão inseridos no contexto de processamento de registros textuais?

   Todos os modelos descritos no Apêndice E, de fundamentação teórica sobre inteligência artificial, são métodos aplicáveis na análise de texto. Contudo, pela literatura não é possível estabelecer o melhor método, é necessário avaliar o problema e realizar experimentações para encontrar a abordagem mais adequada.

2. Como é a utilização do Aprendizado de Máquina para processamento de registros textuais?

   O objetivo de utilização das abordagens da AI na análise de texto, pode ser descrito pelas tarefas de mineração de texto: detecção de tópicos; análise de sentimento; agrupamento; sumarização; suporte na tomada de decisões; e principalmente, classificação.

3. Qual a contribuição do Aprendizado de Máquina, quando utilizado no processamento de textos, para área da Visualização da Informação?

   As abordagens de AI possibilitam estruturar documentos em corpus que podem ser utilizados por técnicas da área de visualização, além disso, a aplicação de métodos de agrupamento e classificação de documentos de texto, permitem a realização da análise visual do corpora.

4. Quais são os conceitos e/ou tecnologias que se relacionam com as áreas de Inteligência Artificial e Aprendizado de Máquina, que possam se inter-relacionar com a área da Visualização da Informação?

   Destaca-se a correlação das áreas AI e visualização em aplicações relacionadas com *Business Intelligence (BI)* e *Visual Analytics*.

B.4.2.1 Diretrizes de Pesquisa

O conhecimento consolidado a partir deste estudo, revela que o problema motivador descrito na Item B.1.1, pode ser abordado utilizando de forma conjunta Classificação de Texto e Processamento de Linguagem Natural, estabelecendo as seguintes diretrizes de pesquisa:

- O problema de classificação de categorias de registros contábeis, pode ser encarado como um problema de classificação hierárquica monorrótulo, com predição opcional em nós folha.
- A identificação de finalidade relacionada com o registro contábil, pode ser abordada com tarefas de NLP ou adaptações de técnicas semânticas como dicionários léxicos ou LDA.

B.4.3 Conclusão da RSL

Este capítulo apresentou um estudo sobre o tema Análise de Texto, com intuito de identificar conceitos, métodos, aplicações e contribuições da área de Aprendizado de Máquina para área de Visualização da Informação. A realização da Revisão Sistemática da Literatura possibilitou a realização da introdução aos tópicos fundamentais da Inteligência Artificial, da Classificação de Texto e do Processamento de Linguagem Natural. A RSL contribuiu para identificação dos processos e procedimentos relacionados com a classificação, auxiliou na descoberta de

métodos/técnicas e ferramentas aplicadas e relacionou finalidades de aplicações com tarefas de TM e NLP. Por fim, esta RSL contribuiu na elucidação de termos e conceitos relacionados aos temas introduzidos, estabelecendo uma base de conhecimento científico fundamental para elaboração do projeto de pesquisa de mestrado. Conclui-se com este estudo, que é viável a aplicação conjunta de métodos e técnicas das áreas Inteligência Artificial e Visualização da Informação, na construção de sistemas inteligência comercial (BI) e análises visuais (*visual analytics*), para descoberta de informações persistidas em registros financeiros e contábeis. Sistemas estes, que possibilitam trabalhar os dados dos registros contábeis, de forma mais amigável e menos burocráticas para o usuário humano.

APÊNDICE C – Protocolo da RSL

Objetivo

Esta revisão tem por objetivo identificar técnicas, métodos, aplicações, usos e contribuições do aprendizado de máquina para área de visualização da informação, ao se trabalhar com dados em registros textuais no formato alfanumérico, considerando qual a contribuição ou correlação do aprendizado de máquina com a visualização da informação textual, sob a perspectiva humana da relação de interação humano-computador.

Questões de pesquisa

Como a inteligência artificial e o aprendizado de máquina podem ser utilizados na construção de ferramentas de visualização de informações textuais?

1. Quais são os métodos e técnicas de Aprendizado de Máquina que estão inseridos no contexto de processamento de registros textuais?
2. Como é a utilização do Aprendizado de Máquina para processamento de registros textuais?
3. Qual a contribuição do Aprendizado de Máquina, quando utilizado no processamento de textos, para área da Visualização da Informação?
4. Quais são os conceitos e/ou tecnologias que se relacionam com as áreas de Inteligência Artificial e Aprendizado de Máquina, que possam se inter-relacionar com a área da Visualização da Informação?

Seleção de fontes

As produções deverão estar em fontes disponíveis na *Web*:

- Em bases científicas cujo acesso seja possível através da conexão da Unesp;
- Em bases indexadas que disponibilizam acesso completo às produções;
- Em bases ou bibliotecas de acesso livre, com produções disponibilizadas diretamente por autores ou instituições

Palavras-chave

"Multi-label", "Natural Language Processing", "text classifier", "textual classification", financial, accounting, accountability

Lista de bases científicas

Google Scholar — https://scholar.google.com.br/

Biblioteca Digital da ACM — https://dl.acm.org

Biblioteca Digital do IEEE — https://ieeexplore.ieee.org

Biblioteca Digital da ScienceDirect — https://www.sciencedirect.com/

Tipos de artigos

- Artigos de estudos primários;
- Artigos de estudos secundários sobre os assuntos de interesse;
- Teses e dissertações com assunto ou objetivo semelhante ao proposto nesta revisão.

Idiomas dos artigos

- Inglês, de forma ampla.
- Português, somente para trabalhos nacionais.

Processo de seleção dos estudos

Será construída uma *string* de busca modelo com a combinação lógica das palavras chave. A partir desta *string* modelo, serão construídas strings personalizadas para cada base de relacionada no no tópico de Fontes. Após a leitura do título e/ou resumo, o principal revisor deverá aplicar os critérios de inclusão e exclusão de trabalhos, encaminhando para o orientador os casos duvidosos ou não previstos nos critérios de inclusão e exclusão.

Estratégia de extração da informação

Trabalhos definitivamente incluídos serão lidos na íntegra e o revisor fará um resumo de cada um deles, destacando os métodos, técnicas e abordagens utilizadas para resolução do problema estudado. Serão preenchidos Formulários de Extração de Dados para cada estudo lido integralmente. Estes formulários devem coletar dados bibliográficos, institucionais, trechos e imagens de destaque.

Sumarização dos resultados

A partir dos dados extraídos nos formulários, será elaborado um relatório técnico com análises quantitativa e qualitativa dos trabalhos.

Critérios de inclusão, exclusão e qualidade

Critérios de inclusão

1. Trabalhos estilo *survey* ou de revisão bibliográfica sobre os assuntos das palavras-chave;
2. Artigos que apresentam métodos ou técnicas do aprendizado de máquina no processamento textual;
3. Artigos que abordam usos ou objetivos da utilização do aprendizado de máquina no processamento textual;
4. Artigos que relacionam ML com a área de visualização da informação;
5. Artigos que abordam a relação humana com o ML;
6. Serão incluídos trabalhos correlatos indicados pelo orientador e/ou membros do grupo de pesquisa;
7. Serão incluídos trabalhos indicados para fundamentação teórica;
8. Serão incluídas indicações de trabalhos estilo *survey* sobre temas de interesse.

Critérios de exclusão

1. Serão excluídos artigos que não estejam disponíveis integralmente na *Web*;
2. Serão excluídos artigos de estudos primários cujo tema não seja relacionado com o processamento de texto;
3. Serão excluídos artigos de estudos secundários cujo data de publicação tenha mais que dez anos;
4. Serão excluídos artigos que não estejam nas línguas de interesse;
5. Artigos de estudos primários com temática não relacionada a esta revisão;
6. Artigos de estudos secundários sobre assuntos diversos aos de interesse;
7. Teses e dissertações com assunto ou objetivo distintos aos de interesse;
8. Serão excluídos artigos por indicação do orientador;
9. Poderão ser excluídos artigos caso o título aparentemente não seja aderente a nenhum assunto de interesse;
10. Poderão ser excluídos artigos que não atendam critérios de qualidade que serão definidos após a 1ª execução de buscas;

Critério de qualidade

1. O trabalho deve ser revisado por pares ou aprovado por banca examinadora.

C.1 Estudos incluídos na busca preliminar

Relação de trabalhos incluídos na RSL.

Referência	Título	Critério inclusão
Amani e Fadlalla (2017)	*Data mining applications in accounting: A review of the literature and organizing framework*	1,5
Das, Mehta e Subramaniam (2017)	*AnnoFin - A hybrid algorithm to annotate financial text*	2,3,5
Fisher, Garnsey e Hughes (2016)	*Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research*	1,5
Giannopoulou e Mitrou (2018)	*Extensive Experimental Evaluation of Self-Organizing Maps for Automatic Classification of a Multi-Class Multi-Label Corpus*	2,3
Guo, Shi e Tu (2016)	*Textual analysis and machine leaning: Crack unstructured data in finance and accounting*	2,3,4,5
Huang (2010)	*Exploring the information contents of risk factors in SEC form 10-K: A multi-label text classification application*	2,3
Huang e Li (2011)	*A multilabel text classification algorithm for labeling risk factors in SEC form 10-K,*	2,3
Kumar e Ravi (2016)	*A survey of the applications of text mining in financial domain*	1,2,3,4,5
Medeiros (2018)	*TagTheWeb: Using Wikipedia Categories to Automatically Categorize Text-Based Resources on The Web*	3,5
Metz (2011)	Abordagens para aprendizado semissupervisionado multirrótulo e hierárquico	2,3,5
Mirończuk e Protasiewicz (2018)	*A recent overview of the state-of-the-art elements of text classification*	1,2,3,5
Mustafi, Mustafi e Sahoo (2020)	*A novel approach to text clustering using genetic algorithm based on the nearest neighbour heuristic*	2,3,4
Pink (2017)	*Slot filling*	3,5
Raza et al. (2019)	*A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews*	2,3,5
Rios e Kavuluru (2018)	*Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces*	2,3,5
Schröder (2018)	*Hierarchical Multiclass Topic Modelling with Prior Knowledge*	6
Song (2009)	*Machine learning for text mining: classification, retrieval and recommendation*	2,3,4,5
Young et al. (2018)	*Recent Trends in Deep Learning Based Natural Language Processing*	1,2,3,5
Zayas et al. (2017)	*Getting ready for data analytics of electric power distribution systems*	6
Zheng e Zheng (2019)	*A Hybrid Bidirectional Recurrent Convolutional Neural Network Attention-Based Model for Text Classification*	2

Fonte: Produzida pelo autor.

C.2 Estudos incluídos na busca em bases

Relação de trabalhos incluídos na atualização da RSL.

Referência	Título	Critério inclusão
Bi et al. (2020)	*Mining Knowledge within Categories in Global and Local Fashion for Multi-Label Text Classification*	2,3
Bittencourt, Silva e Almeida (2020)	*ML-MDLText: An efficient and lightweight multilabel text classifier with incremental learning*	2,3
Cheng, Nazarian e Bogdan (2020)	*VRoC: Variational Autoencoder-aided Multi-task Rumor Classifier Based on Text*	3
Gong, Shi e Niu (2019)	*Hierarchical Text-Label Integrated Attention Network for Document Classification*	2,3
Huang et al. (2019)	*Hierarchical Multi-label Text Classification: An Attention-based Recurrent Network Approach*	2,3
Ibrahim et al. (2021)	*GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification*	2,3,5
Li et al. (2016)	*Multi-label maximum entropy model for social emotion classification over short text*	2,3
Liu et al. (2017)	*Deep Learning for Extreme Multi-label Text Classification*	2,3
Liu et al. (2021)	*Multi-label text classification via joint learning from label embedding and label correlation*	2,3
Ma et al. (2022)	*Hybrid embedding-based text representation for hierarchical multi-label text classification*	2,3
Maltoudoglou et al. (2022)	*Well-calibrated confidence measures for multi-label text classification with a large number of labels*	2,3,5
Santos e Merschmann (2020)	*Metalearning Applied to Multi-label Text Classification*	2,3,5
Tao, Cui e Wenjun (2018)	*A Multi-Label Text Classification Method Based on Labels Vector Fusion*	2,3,5
Wang e Tan (2021)	*Label-Based Convolutional Neural Network for Text Classification*	2,3
Xiang e Zheng (2018)	*Multi-Label Emotion Classification for Imbalanced Chinese Corpus Based on CNN*	2,3
Xiao et al. (2021)	*History-based attention in Seq2Seq model for multi-label text classification*	2,3

Fonte: Produzida pelo autor.

C.3 Estudos excluídos

Relação de trabalhos excluídos da RSL.

Referência	Título	Critério exclusão
Eshima, Imai e Sasaki (2020)	*Keyword Assisted Topic Models*	8 (PRÉ-PRINT)
Eshima (2020)	*Supplementary Appendix for "Keyword Assisted Topic Models"*	DUPLICADO

Fonte: Produzida pelo autor.

APÊNDICE D – Fundamentos da análise de texto

A Análise de Texto, mais conhecida na literatura como Mineração de Texto (GUO; SHI; TU, 2016), possui estreita relação com a área de Mineração de Dados. Desta forma, este capítulo discorre sobre conceitos fundamentais para melhor entendimento dos assuntos destacados no título deste trabalho. Os conceitos abordados neste capítulo começam evidenciando a relação da Mineração de Dados com a Mineração de Texto, culminando nos tópicos de interesse desta monografia, Classificação de Texto e Processamento de Linguagem Natural.

D.1 *Data Mining*

A Mineração de Dados (DM) é um processo multidimensional para descoberta do conhecimento, baseado na identificação de padrões interessantes presentes em grandes volumes de dados. As principais dimensões da mineração são os dados, o conhecimento, as tecnologias envolvidas e as aplicações. Neste processo, normalmente estão envolvidas as tarefas de limpeza de dados, integração de dados, seleção de dados, transformação de dados, descoberta de padrões, avaliação de padrões e apresentação/visualização de conhecimento (HAN; PEI; KAMBER, 2011).

Amani e Fadlalla (2017) comentam que a utilização da mineração de dados na descoberta do conhecimento é aplicada com três principais objetivos: descrição, predição e prescrição.

- Descrição — Busca o entendimento e explicação do passado (o que aconteceu), através da busca de padrões humanamente interpretáveis que descrevam os dados (AMANI; FADLALLA, 2017).
- Predição — Utiliza o conhecimento prévio para inferir o desconhecido (o que poderia acontecer), utilizando-se de campos conhecidos do banco de dados, para prever valores de outros campos ou informações de interesse (AMANI; FADLALLA, 2017).
- Prescrição — Abordagem que busca recomendar a melhor solução para um problema em questão (o que deveria acontecer). (AMANI; FADLALLA, 2017).

Sendo que esses objetivos podem ser alcançados utilizando muitas Tarefas de Mineração de Dados como: classificação, clusterização, predição, detecção de anomalia, otimização e visualização (AMANI; FADLALLA, 2017).

D.2 *Text Mining*

Song (2009) aborda a Mineração de Texto (TM) como um ramo de pesquisa da grande área do conhecimento denominada Gestão da Informação, descrevendo que este tipo de mineração

significa o processo de descoberta de padrões úteis, estruturas e outras informações valiosas de textos em linguagem natural e não estruturados. Ou seja, a Mineração de Texto, que também pode ser referida como Análise de Texto (GUO; SHI; TU, 2016), descreve a área de específica de DM para o processamento de dados em formato textual.

D.2.1 Finalidades de aplicação da análise de texto

Das muitas finalidades de aplicação, o uso da TM frequentemente tem seu propósito relacionado com: agrupamento de documentos, classificação de documentos, resumo de texto, análise de sentimento, análise de rede social, detecção de tópicos, classificação de página da *Web*, identificação do autor, detecção de plágio, análise de patentes, detecção de *phishing*/*spam*/*malware*, tomada de decisão financeira, etc. Sendo estas aplicações normalmente relacionadas com uma das cinco principais categorias de Tarefas de Mineração de Texto: classificação, agrupamento, mineração de regras de associação, resumo de texto e detecção/identificação de tópicos (KUMAR; RAVI, 2016).

Considerando os objetivos descritos por Amani e Fadlalla (2017) como objetivos gerais de DM, bem como TM sendo uma especificação desta, é possível assumir que as tarefas listadas por Kumar e Ravi (2016) são objetivos específicos da Mineração de Texto. Desta forma, a Tabela 21 apresenta a relação de objetivos e finalidades de aplicação de TM.

Tabela 21 – Finalidades de aplicação da Mineração de Texto.

Objetivo geral	Objetivo específico	Finalidade de aplicação*
Descrição	Classificação	Classificação de documentos Categorização de páginas *Web*
	Análise de sentimento	Identificação de humor Detecção de depressão Identificação de tom do discurso
Predição	Detecção de tópicos	Indexação de documentos Resolução de ambiguidade de palavras
	Agrupamento	Segmentação de clientes Detecção de anomalias Sistemas de recomendação
	Sumarização	Resumo de texto Criação de tesauros
Prescrição	Tomada de decisões	Recomendação de compra e venda de ações Indicação de periódico para publicação

∗ Apenas alguns exemplos, mas não limitadas à estas aplicações.

Fonte: Produzida pelo autor.

D.2.2 Abordagens da análise de texto

Conforme descrito por Guo, Shi e Tu (2016), é possível distinguir abordagens de análise de texto em três categorias: lexical, semântica e *Machine Learning*.

No contexto do estudo dos autores, a abordagem lexical se refere a técnicas relacionadas com medidas de legibilidade de texto como *Fog Index* ou, refere-se a técnicas baseados em

dicionários léxicos como *Harvard General Inquirer (Harvard GI)* ou *Loughran and Mcdonald dictionary (LM dictionary)*. Mas os autores também citam o trabalho de Ingram e Frazier (1980), como sendo o primeiro trabalho a introduzir a análise de conteúdo textual, utilizando-se da contagem de frequência de palavras-chave para realizar esta abordagem lexical.

Guo, Shi e Tu (2016) explicam ainda que a abordagem semântica dedica-se à extração de conteúdo conceitual em documentos de texto, buscando também a identificação de relações entre documentos. A última abordagem, *Machine Learning*, relaciona-se com tarefas específicas de análise de texto.

A possibilidade de analisar dados não estruturados utilizando técnicas de TM possibilita ao usuário uma melhor tomada de decisão. Apesar da Mineração de Texto remontar da década de 1960, ela tornou-se popular apenas na última década do século XX, a partir da introdução de algoritmos de *Machine Learning* para a realização de tarefas de mineração de texto, o que reduziu drasticamente a necessidade de intervenção humana, bem como o tempo para análise do texto (KUMAR; RAVI, 2016).

D.3 Classificação de Texto

Numa definição ampla, a Classificação de Texto é o processo de atribuir categorias a documentos de texto. Este conceito de também pode ser encontrada na literatura como Categorização de Texto (SONG, 2009), (ZHENG; ZHENG, 2019), (KUMAR; RAVI, 2016), (MEDEIROS, 2018) e (RAZA et al., 2019). Apesar da conceitualização ampla ser bastante simples, esta definição varia de acordo com o modo operante e com o problema de classificação abordado.

No quesito modo operante, a classificação pode ser automática ou manual. Quando realizada manualmente, denomina-se Classificação Manual de Texto e depende de humanos com expertise no domínio de aplicação para atribuir rótulos a cada amostra de texto, um trabalho árduo e demorado, tornando-se inviável para corporas com grandes volumes de documentos (SONG, 2009) e (FISHER; GARNSEY; HUGHES, 2016).

A Classificação Automática de Texto é a mais proeminente dentre todas as Tarefas da Mineração de Texto. No estudo de Amani e Fadlalla (2017), 67% dos trabalhos estavam relacionados com classificação. E Song (2009) afirma que a tarefa de classificação é uma das mais importantes áreas de pesquisa, nos campos da Mineração de Texto e *Machine Learning*.

Considerando o foco deste trabalho, a abordagem automática será denominada simplesmente por Classificação de Texto. Quando necessário, a outra abordagem será identificada explicitamente como Classificação Manual de Texto.

D.3.1 Problemas de classificação

Com relação ao problema de classificação, a nomenclatura dos tipos de problemas é baseada em características específicas de cada problema. Sendo que um mesmo problema pode ser enquadrado em mais de um tipo. Nota-se na literatura a ocorrência de diferenças conceituais na utilização de termos da nomenclatura, que se referem a tipos de problemas de classificação.

Aparentemente as diferenças de língua materna e língua escrita dos trabalhos contribui tanto para variações como para confusões. Os termos em inglês *multi-class* e *multi-label*, foram traduzidos neste trabalho respectivamente para multiclasse e multirrótulo. Pois é desta forma que foram aplicados em grande parte dos estudos referenciados: Song (2009), Huang (2010), Huang e Li (2011), Giannopoulou e Mitrou (2018), Mirończuk e Protasiewicz (2018), Raza et al. (2019) e, Zheng e Zheng (2019).

- Classificação binária — Modelo de problema onde o conjunto de soluções é de ordem binária. O conjunto de classes possíveis de serem atribuídas às amostras consiste em apenas duas opções (MIROŃCZUK; PROTASIEWICZ, 2018).

  Exemplos de classes para problemas binários: [0,1], [verdadeiro, falso] e [receita, despesa].

- Classificação multiclasse — Neste modelo, o conjunto de soluções tem um tamanho finito variando de 3 até n (MIROŃCZUK; PROTASIEWICZ, 2018).

  Exemplos de classes para problema multiclasse: [auxílios, compras, diárias, transportes].

- Classificação monorrótulo — Abordagem relacionada aos problemas multiclasse, normalmente com categorias mutuamente excludentes. É um problema onde o domínio da aplicação aceita a atribuição de uma e somente uma categoria para cada amostra observada (METZ, 2011).

  Exemplos de categorias para problema monorrótulo: [pessoa, veículo, planta].

- Classificação multirrótulo — Abordagem também relacionada aos problemas multiclasse. Mas neste caso aceita-se a atribuição de várias categorias para cada amostra observada (METZ, 2011).

  Exemplo: Utilizando as categorias [ensino, pesquisa, extensão], uma amostra pode ser categorizada como atividade relacionada ao ensino e pesquisa.

- Classificação plana — Problema de classificação onde o conjunto de classes possíveis de serem atribuídas às amostras, não apresenta relação de generalização ou especialização. Ou seja, o conjunto de classes é formado por categorias independentes umas das outras (METZ, 2011).

  Exemplos de categorias para classificação plana: [ensino, pesquisa, extensão].

- Classificação hierárquica — Modelo que apresenta relações de especialização ou generalização entre as classes do conjunto de categorias possíveis (METZ, 2011).

  Exemplos de classes hierárquicas: [despesas, despesas.auxilio, despesas.auxilio.pesquisador, despesas.auxilio.estudante, despesas.auxilio.colaboradoreventual].

- *Few-Shot* e *Zero-Shot* — Abordagens de classificação que visam categorizar amostras observadas na avaliação, utilizando classes que estão pouco presentes (*Few-Shot*) ou completamente ausentes (*Zero-Shot*) do conjunto de dados de treinamento (RIOS; KAVULURU, 2018).

Exemplo: Considerando o conjunto de soluções como sendo todas as doenças identificadas, é possível que enfermidades extremamente raras não sejam encontradas no conjunto de treinamento, mas sejam encontradas posteriormente em conjuntos de dados de teste, avaliação ou produção.

Nota-se que algumas das nomenclaturas expostas possuem características opostas: plana/hierárquica; binária/multiclasse; e, monorrótulo/multirrótulo. De modo que esta oposição de características incute uma restrição semântica na identificação de tipos de problema. Sendo assim, observa-se na literatura a utilização de nomenclaturas que possuem sentido implícito, por exemplo: classificação binária (é uma classificação plana), classificação multirrótulo (normalmente é multiclasse) e classificação hierárquica (é uma classificação multiclasse).

D.3.2 Abordagens para problemas de classificação

Problemas de classificação com diferentes características, apresentam diferentes abordagens de enfrentamento. Sendo estas abordagens relacionadas diretamente com a estrutura do problema ou com a quantidade de rótulos que poderá ser atribuído às amostras.

◆ Problemas de classificação hierárquica

Segundo Metz (2011) o problema da classificação hierárquica de texto, pode ser desenvolvido utilizando-se duas abordagens de classificação:

- *Mandatory Leaf Node Prediction (MLN)* — A Predição Obrigatória em Nós-Folha, é a abordagem que considera somente classes alocadas em nós terminais, como elegíveis para atribuição às amostras.
- *Non-mandatory Leaf Node Prediction (NMLN)* – A Predição Opcional em Nós-Folha, permite que qualquer classe da estrutura hierárquica seja atribuída às amostras observadas.

◆ Problemas monorrótulo

Schröder (2018) destaca que quando as classes do conjunto de categorias são mutualmente excludentes, duas abordagens podem ser adotadas para realização da classificação de amostras:

- *One-vs-All* — Um-contra-Todos é um método que cria K classificadores, um para cada classe do conjunto, então as amostras são submetidas a todos estes classificadores. Aquele que obtiver a menor diferença (maior semelhança) é a classe que atribuirá rótulo à amostra.
- *One-vs-One* — Na abordagem Um-contra-Um o problema multiclasse é transformado em vários problemas de classificações binárias. Considerando um conjunto de classes K, serão gerados $K(K-1)/2$ classificadores binários para realização de comparações de duas em duas classes. O classificador que ganhar mais embates 1x1 será a classe atribuída à amostra.

◆ Problemas multirrótulo

Quando as classes do conjunto de categorias não são mutualmente excludentes, ou seja, a amostra pode ser rotulada com várias categorias, Metz (2011) e Schröder (2018) apresentam três abordagens possíveis:

- *Binary Relevance (BR)* — Na Relevância Binária, constrói-se um classificador discriminante (é ou não é) para cada rótulo possível. A amostra será analisada por todos estes classificadores e serão atribuídos os rótulos considerados relevantes.
- *Label Powerset (LP)* — Esta abordagem consiste na transformação do problema multiclasse-multirrótulo em problema multiclasse-monorrótulo. Criando-se novos rótulos que são compostos pelas combinações de rótulos possíveis no conjunto de treinamento.
- *Stacking* — A ideia do Empilhamento é a utilização de metaclassificadores, onde os de mais alto nível utilizam como entrada a saída gerada pelos metaclassificadores base.

D.3.3 Processo de classificação de texto

Kumar e Ravi (2016) exibem uma visão extremamente simplificada do processo de mineração de texto, considerando apenas duas fases: pré-processamento do texto e extração do conhecimento. Em Mirończuk e Protasiewicz (2018), os autores apresentam uma descrição do processo de classificação de texto realizada em cinco etapas: (1) pré-processamento de documento; (2) modelagem de documento; (3) seleção e projeção de características; (4) aplicação de métodos de aprendizado de máquina; e, (5) indicadores de qualidade e métodos de avaliação. Em seguida os autores diferenciam seu trabalho propondo uma visão mais sofisticada e detalhada do processo, considerando seis etapas: (1) aquisição de dados; (2) análise de dados e rotulação; (3) construção de características e ponderação; (4) seleção e/ou projeção de características; (5) treinamento do modelo; e, (6) avaliação da solução.

Considerando as descrições de processo realizadas em (KUMAR; RAVI, 2016) e (MIROŃCZUK; PROTASIEWICZ, 2018), agrupa-se as etapas do processo de classificação em três grandes fases: pré-processamento; treinamento; e, validação.

◆ Pré-processamento

Apesar de diferentes perspectivas sobre o processo de classificação, é consenso que a fase de pré-processamento é fundamental para qualquer solução de Mineração de Texto, pois ações realizadas neste ponto do processo influenciam diretamente nos resultados finais (KUMAR; RAVI, 2016), (MEDEIROS, 2018), (MIROŃCZUK; PROTASIEWICZ, 2018) e (RAZA et al., 2019).

Na fase de pré-processamento são realizadas análises morfológicas, semânticas e sintáticas das palavras que compõe o texto (SANTOS; MERSCHMANN, 2020). A realização deste conjunto de análises culmina em duas grandes ações que são realizadas nesta fase: (i) modelagem de documento e (ii) seleção de características e tratamento de dimensionalidade.

- Modelagem de documento — A modelagem de documento tem como objetivo a transformação do documento de texto original, em um modelo que seja adequado para ser processado por algoritmos computacionais (MIROŃCZUK; PROTASIEWICZ, 2018).

  Procedimentos executados:

  a) *Case folding* – Padronização de texto em caixa baixa ou alta (frequentemente em baixa).

  b) *Cleaning* – Remoção de caracteres especiais, dígitos, sinais de pontuação e acentuação.

  c) *Stopwords removal* – Remoção de palavras irrelevantes como artigos e preposições.

  d) *Length filtering* – Remoção de palavras pequenas (normalmente menor que 3 caracteres).

  e) *Tokenization* – Transformação de palavras em tokens *n-grams*, onde n indica a quantidade máxima de palavras que formam um token (*1-gram* é o mais comum, mas encontra-se usos de *2-grams* e *3-grams*).

  f) *Stemming* – Redução de palavras ao seu radical.

  g) *Lemmatisation* – Alteração de formas flexionadas de palavras, para uma forma padrão.

  h) *Document representation* – Construção da matriz de termos do documento e definição de modelo de representação (vetorial, grafos, PoS, etc.).

- Extração de características e tratamento de dimensionalidade — Parte mais importante do pré-processamento, consequentemente também é considera parte mais importante da Mineração de Texto (KUMAR; RAVI, 2016). Na literatura existem muitos estudos que concentram-se apenas neste tópico (MIROŃCZUK; PROTASIEWICZ, 2018).

  Procedimentos executados:

  a) *Feature selection* – Seleção de tokens que representam características do documento.

  b) *Feature weighting* – Ponderação de características com atribuição de peso aos tokens selecionados.

  c) *Feature projection* – Projeção de características em modelo de menor dimensão.

  d) *Feature scaling* – Transformação dimensional do espaço de características.

  e) *Instance selection* – Seleção de amostras para redução do espaço de instâncias.

◆ Treinamento do modelo

Nesta fase, o modelo computacional escolhido passa pela indução de treinamento (MEDEIROS, 2018).

Procedimentos executados:

a) Particionamento — Definição dos conjuntos de dados para treinamento, teste e avaliação.

b) Balanceamento amostral — Correções e ajustes na distribuição de amostras por conjuntos.

c) Treinamento do algoritmo — Apresentação do conjunto de treinamento ao algoritmo escolhido.

d) Testagem e tunagem — Ciclo de testes e ajustes de parâmetros e hiperparâmetros para até atingir melhor desempenho possível.

e) Construção do modelo treinado — Persistência do modelo com suas informações de treinamento (o conhecimento aprendido).

◆ Validação e avaliação

Fase dedicada aos testes qualitativos do modelo treinado (MIROŃCZUK; PROTASIEWICZ, 2018).

Procedimentos executados:

a) Escolha de métricas — Determinação de índices de avaliação (*Accuracy*, *Precision*, *Recall*, *F-Measure*, etc.).

b) Escolha de método — Determinação do método de execução de testes avaliativos (*leave-one-out*, *k-Fold cross-validation*, etc.).

c) Validação — Execução do modelo de avaliação definido, cômputo e sumarização de resultados.

d) Avaliação — Cálculo dos valores de índices de avaliação, comparação com modelo de base.

D.3.4 Considerações sobre classificação

Apesar de vários procedimentos terem sido relatados nas etapas do processo de classificação, não há a obrigatoriedade de todos serem executados. É possível em alguns casos a escolha de determinado procedimento em detrimento de outro, como exposto em Raza et al. (2019), onde os autores preferiram utilizar *Lemmatisation* ao invés de *Stemming*.

Considerando o procedimento de extração de características e tratamento de dimensionalidade, *Term Frequency – Inverse Document Frequency (TF–IDF)* é o método mais popular para calcular a importância de um termo. Baseia-se na premissa de que, se uma palavra (termo) é frequente em um conjunto de documentos de mesma classe, mas pouco frequente no demais documentos do corpora, considera-se assim, que este é um termo altamente discriminatório para a classe (MUSTAFI; MUSTAFI; SAHOO, 2020).

A análise semântica de texto refere-se à compreensão do significado das palavras no texto, seja pela utilização de dicionários ou pela extração de seus contextos, sendo que a detecção de palavras-chave é uma das possíveis técnicas aplicáveis para esta finalidade (SANTOS; MERSCHMANN, 2020). *Latent Dirichlet Allocation (LDA)* também é uma das técnicas utilizadas para abordagem semântica da análise de texto, sendo utilizado frequentemente para extração de características e tratamento de dimensionalidade (GUO; SHI; TU, 2016). LDA é um modelo

probabilístico generativo para coleção de dados discretos como corpora de texto, considerado também um método de aprendizagem não supervisionada. É um modelo naturalmente aplicado na identificação de tópicos latentes em coleções de documentos, sendo possível estendê-lo para utilização com outras finalidades de aplicação (SCHRÖDER, 2018).

D.4 *Natural Language Processing*

O Processamento de Linguagem Natural (NLP, do inglês, *Natural Language Processing*) é a área do conhecimento que estuda técnicas computacionais para representação e análise automática da linguagem humana. A implementação destas técnicas frequentemente está relacionada com abordagens de inteligência artificial como *Machine Learning* e *Deep Learning* (YOUNG et al., 2018).

D.4.1 Tarefas de Processamento de Linguagem Natural

Young et al. (2018) descrevem em seu estudo as principais tarefas de NLP:

- *Semantic Role Labeling (SRL)* — A Rotulagem de Função Semântica é a identificação de estrutura linguística, onde para cada verbo alvo busca-se encontrar demais constituintes da oração que estabelecem função semântica.
- *Part-of-Speech (PoS)* — Parte-da-Fala realiza a análise sintática da linguagem para identificação da função gramatical das palavras.
- *Named Entity Recognition (NER)* — Reconhecimento de Entidades Nomeadas é a tarefa de identificação de palavras relacionadas à pessoas, locais, organizações e entidades diversas.
- *Information Retrieval (IR)* — Recuperação da Informação é a tarefa de recuperação de documentos baseada em *queries* de busca.
- *Speech recognition* — Reconhecimento da fala.
- *Machine translation* — Realização de tradução automática por computador.
- *Question & Answering (Q&A)* — Sistemas de perguntas e respostas.
- *Summarization* — Sumarização, criação automática de resumos de textos.
- *Sentence classification* — Classificação de sentenças é a atribuição de classe à sentenças de texto.
- *Dialogue systems* — Sistemas de diálogos, que podem ser baseados na geração de linguagem natural ou na recuperação das respostas mais adequadas dentre as cadastradas no repositório.

Devido a própria característica da linguagem natural, o NLP é amplamente empregado em análises semânticas de textos, apresentando tarefas relacionadas com este propósito (CHENG; NAZARIAN; BOGDAN, 2020), (FISHER; GARNSEY; HUGHES, 2016), (KUMAR; RAVI, 2016) e (YOUNG et al., 2018).

D.4.2 A Maldição da Dimensionalidade

Young et al. (2018) comentam que processamento de linguagem natural estatístico, realizado com técnicas simples de *Machine Learning*, foi a primeira opção para modelar as complexas tarefas de NLP, mas essas técnicas sofrem muito com a temida Maldição da Dimensionalidade (*The Curse of Dimensionality*). Metz (2011) explica que esta maldição ocorre quando os dados são representados por conjuntos extremamente grandes de características, tornando inviável a solução devido ao alto custo de tempo computacional.

Para evitar ou solucionar este problema, recomenda-se a aplicação de técnicas de redução de dimensionalidade. Processo no qual as características mais relevantes do texto são selecionadas para representá-lo. Na primeira etapa do processo, estas características são extraídas e ponderadas com técnicas adequadas para tal. Na última etapa do processo, as características representativas são projetadas em um espaço de característica de menor dimensão do que o original (KUMAR; RAVI, 2016) e (MIROŃCZUK; PROTASIEWICZ, 2018).

D.4.3 Considerações sobre NLP

A utilização de métodos mais sofisticados, que podem ser categorizados como métodos *Deep Learning* ou *Artificial Intelligence*, lidam melhor com dados em alta dimensionalidade, apresentando melhor desempenho nas tarefas de NLP (YOUNG et al., 2018).

D.5 Considerações finais sobre a Análise de Texto

Song (2009) comenta sobre a vasta abrangência do tema Mineração de Texto e opta por abordar aspectos que considera mais relevante para sua tese. Neste sentido, a discussão detalhada neste capítulo foi limitada aos dois tópicos diretamente relacionados com o problema relatado no Capítulo 1: Classificação de Texto e Processamento de Linguagem Natural.

> *The research issues of text mining have been studied for decades, by researchers from different research areas including applied mathematics, statistics, machine learning, natural language processing and etc. Apparently, we are unable to cover all sub areas of text mining and thus we will focus on four important areas in this thesis: text classification, text retrieval, text recommendation and topic discovery (SONG, 2009).*

APÊNDICE E – Fundamentos da inteligência artificial

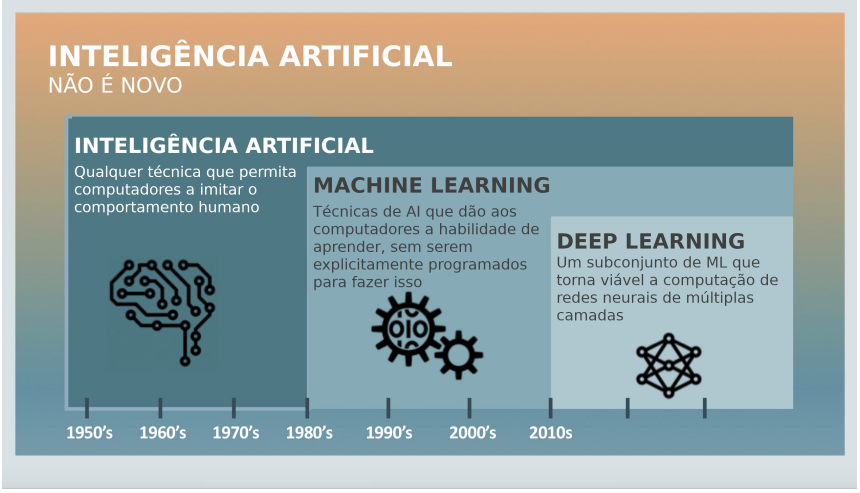
Capítulo dedicado à apresentação de conceitos, abordagens, métodos e técnicas relacionadas com inteligência artificial. Os modelos computacionais descritos aqui, são algumas das ferramentas que possibilitam a implementação da Classificação de Texto e do Processamento de Linguagem Natural.

E.1 Perspectiva sobre Inteligência Artificial

A Inteligência Artificial (AI) será abordada nesta monografia como uma área do conhecimento, historicamente consolidada (GOMES, 2010). Apesar da definição conceitual de AI não estar no escopo deste trabalho, julga-se adequado estabelecer uma perspectiva sobre esta área do conhecimento suas relações.

IEEE e ACM são instituições referência para pesquisas na área da Computação, mesmo estas instituições apresentam diferentes visões sobre o tema Inteligência Artificial (AI). O ACM CCS, sistema de classificação conceitual da ACM (2012), indica *Artificial Intelligence* e *Machine Learning* como áreas irmãs, ambas diretamente subordinadas de *Computing Methodologies*. Ao passo que a Taxonomia do IEEE (2017), considera *Machine Learning* como área filha de *Artificial Intelligence*. A perspectiva adotada nesta monografia coincide com a visão do IEEE. A Figura 29 ilustra esta perspectiva.

Figura 29 – Relação adotada para as áreas de AI, ML e DL.



Fonte: Adaptado de Jeffcock (2018).

E.2 *Machine Learning*

Aprendizado de Máquina (ML) é um campo de pesquisa da área da computação, relacionado com a teoria da aprendizagem computacional, com foco no estudo de métodos para simular o aprendizado humano em computadores, por meio de regras de aprendizagem e generalização baseada em exemplos. Tem origem nos meados do século XX e foi popularizado por Samuel (1959) com a publicação de estudos dedicados ao ensinamento de xadrez para computadores (FRIEDBERG, 1958; MICHIE, 1968).

E.2.1 Paradigmas de aprendizagem

Os métodos de treinamento, ou algorítimo de aprendizagem, podem ser classificados de acordo com a abordagem utilizada no treinamento.

- Aprendizagem supervisionada — Processo que treina uma função de aprendizagem utilizando exemplos de dados, onde todas amostras tem seus valores de entrada e saída conhecidos (MIROŃCZUK; PROTASIEWICZ, 2018).
- Aprendizagem não supervisionada — Ao contrário do aprendizado supervisionado, neste caso os exemplos não tem valores conhecidos (FISHER; GARNSEY; HUGHES, 2016).
- Aprendizagem semissupervisionada — Abordagem onde uma pequena parte das amostras de treinamento tem valores conhecidos, as quais são utilizadas para inferir valores das demais amostras do conjunto de dados de treinamento. Esta abordagem também pode ser encontrada na literatura como: aprendizagem a partir dos dados rotulados e não rotulados, aprendizagem transdutiva, co-treinamento e autotreinamento (MIROŃCZUK; PROTASIEWICZ, 2018).
- Aprendizagem por transferência — Também é conhecido como transferência indutiva ou transferência de conhecimento entre domínios (MIROŃCZUK; PROTASIEWICZ, 2018), refere-se à abordagem onde um modelo é treinado em um domínio e testado em outro domínio de conhecimento (KUMAR; RAVI, 2016).
- Aprendizado por múltiplas visões — Paradigma de aprendizado que utiliza diferentes espaços de características para modelar diferentes visões de um mesmo problema, considerando o conjunto de todas as visões para melhorar o desempenho da generalização. Também pode ser referenciado na literatura como: espaços de características diversificado ou, fusão/integração de dados de múltiplos espaços de características (MIROŃCZUK; PROTASIEWICZ, 2018).
- Aprendizagem por conjunto — Paradigma onde há a utilização de diferentes métodos de treinamento, sendo o resultado da aprendizagem decidido em conjunto pelo comitê de aprendizagem. De certa forma, pode ser considerado um caso especial de aprendizado por múltiplas visões (MIROŃCZUK; PROTASIEWICZ, 2018).

E.2.2 Modelos e Métodos de *Machine Learning*

De uma forma geral, diferentes implementações de algorítimos (métodos) de aprendizagem são realizadas baseadas em um modelo conceitual paramétrico ou não-paramétrico. Modelos baseados em lógica, modelos probabilísticos, redes neurais e modelos baseados em instâncias, são alguns exemplos destes modelos conceituais (RAZA et al., 2019).

- *K-Nearest Neighbour (k-NN)* — O algoritmo dos k-Vizinho(s) Mais Próximo(s) é conhecido como aprendiz preguiçoso, porque efetivamente não realiza nenhuma aprendizagem nem função discriminativa. Este método é um modelo baseado em instâncias que memoriza os dados de treinamento e no momento da testagem verifica a distância da amostra testada em relação às k-amostras memorizadas. O valor de k determina a quantidade de vizinhos considerados na comparação. Normalmente utiliza distância euclidiana, distância de Manhattan ou distância de Minkowski como métricas para calcular a distância entre amostras (KUMAR; RAVI, 2016) e (RAZA et al., 2019).

- *Naive Bayes (NB)* — Método probabilístico baseado no Teorema de Bayes, considerando a independência das variáveis do espaço de características. É muito utilizado devido suas características de facilidade de implementação, velocidade de treinamento e razoável capacidade de previsão (KUMAR; RAVI, 2016; RAZA et al., 2019).

- *Rocchio* — Algoritmo de classificação que tem como ideia central a identificação amostas como relevantes ou não para determinada classe. Cada amostra é representada como um vetor de características em seu modelo de espaço vetorial, sendo que o conjunto de amostras relevantes formam um protótipo amostral (ou meta-amostra) que o representam o centroide da classe. A testagem é realizada pela comparação de distâncias utilizando métricas como as descritas no método k-NN (RAZA et al., 2019).

- *Perceptron* — Algoritmo mais simples dos modelos de rede neural. A versão *single layer* possui apenas uma camada que age como neurônio único de classificação binária. (RAZA et al., 2019). Na versão *multilayer* constam três camadas (entrada, saída e oculta) que formam redes neurais de alimentação sequencial, onde cada camada é alimentada por informações da camada anterior (KUMAR; RAVI, 2016).

- *Linear Regression / Ridge* — O algoritmo de regressão linear busca a identificação da relação de uma variável discreta denominada dependente, com uma ou mais variáveis denominadas independentes. Ou seja, a regressão procura inferir a variável dependente utilizando a(s) variável(is) independente(s). A relação entre as variáveis pode ser calculada utilizando o popular método de estimativa de mínimos quadrados, onde a minimização do erro médio quadrático equivale a maximização de semelhança (RAZA et al., 2019). *Ridge* é a versão regularizada do algoritmo de regressão linear (KUMAR; RAVI, 2016).

- *Stochastic Gradient Descent (SGD)* — Uma variação do algoritmo de otimização chamado Gradiente Descendente, que é utilizado para calcular parâmetros que minimizem a função de custo. Esta variação é apropriada para grandes quantidades de dados, pois vai atualizando seus coeficientes (parâmetros) a cada instância de treinamento (RAZA et al., 2019).

- *Logistic Regression (LR)* — O núcleo da Regressão Logística é a função logística, uma função sigmoidal cujo gráfico apresenta uma curva em forma de S, permitindo mapear qualquer número de valor real dentro do intervalo entre 0 e 1. A função *SoftMax* é uma generalização da função sigmoidal que também pode ser utilizada na implementação deste algoritmo. A Regressão Logística utiliza o conjunto de treinamento para estimar os parâmetros (coeficientes) da função que possibilitam a máxima verossimilhança (RAZA et al., 2019).

- *Decision Tree (DT)* — Uma Árvore de Decisão é uma abordagem de aprendizagem supervisionada, com características de modelagem preditiva e não paramétrica. O modelo representa uma árvore binária do espaço de entrada, sendo cada nó da árvore relacionado com uma característica e cada nó folha representa uma classe (RAZA et al., 2019).

- *Classification And Regression Tree (CART)* — Enquanto as DTs focam basicamente em problemas de classificação, CARTs são utilizadas para tanto para classificação quanto para regressão (KUMAR; RAVI, 2016).

- *Support Vector Machine (SVM)* — Método proposto inicialmente por Vapnik (1995) onde a construção de hiperplanos permite a separação de amostras no espaço de características. Os hiperplanos são criados com base numa função núcleo (*kernel function*), de modo que diferentes implementações com diferentes funções, podem ser denominadas de acordo com sua variação de função (linear-SVM, polinomial-SVM, radial-SVM, sigmoidal-SVM, etc.). A escolha do *kernel* que determina o tipo de separação possível pelo hiperplano (KUMAR; RAVI, 2016; RAZA et al., 2019).

- *Passive Aggressive (PA)* — Similar ao SVM, o algoritmo Passivo Agressivo também é um método baseado no modelo de margem (fronteira ou limite) de decisão. Esta técnica realiza atualização do modelo aprendido conforme processa os exemplos de treinamento. O classificador é atualizado com as restrições, de modo que o novo classificador seja semelhante ao atual (perturbação mínima) e atinja ao menos uma margem unitária (precisão máxima de predição) no exemplo mais recente (RAZA et al., 2019).

- *Gaussian Process (GP)* — O Processo Gaussiano é um processo estocástico, onde um conjunto de variáveis formam uma distribuição gaussiana multivariada, especificada por uma função média e função de covariância. Esta técnica é um modelo não paramétrico que tem mostrado melhor desempenho do que outros métodos de aprendizagem como SVM e k-NN na tarefa de classificação. Modelos de GP têm sido utilizados para aprendizagem bayesiana aproximada, com duas aplicações bem-sucedidas: regressão e classificação (SONG, 2009).

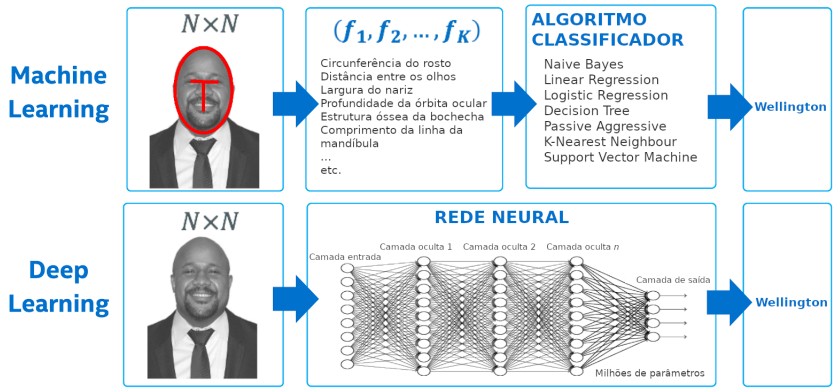
- *Group Method of Data Handling (GMDH)* — Modelo considerado a primeira arquitetura de rede neural de aprendizado profundo. Utilizado para modelar problemas complexos, funciona com base na construção de termos polinomiais. A saída da rede depende da combinação polinomial das entradas, sendo que o número de camadas da rede é determinado por um componente genético (KUMAR; RAVI, 2016).

E.3 *Deep Learning*

O Aprendizado Profundo de Máquina (DL) utiliza um modelo de rede neural que consiste numa enorme quantidade de camadas, permitindo o aprendizado automático de representação de características em seus vários níveis (YOUNG et al., 2018).

A Figura 30 exemplifica a diferença no processo de classificação de imagem[1] realizado por *Machine Learning* e *Deep Learning*, onde nota-se que a etapa de extração de características efetuada na abordagem clássica de ML, é suprimida pela complexa rede neural do modelo DL.

Figura 30 – Comparação *Machine Learning* e *Deep Learning*.



Fonte: Adaptado de Robins (2020).

Em seu estudo sobre aplicações de *Text Mining* no domínio do mercado financeiro, Kumar e Ravi (2016) destacam como direções futuras de pesquisa a utilização de DL, por este modelo ser potencialmente útil em lidar com a alta dimensionalidade do espaço de características nos corpus textuais do domínio. Considerando necessário integrar a abordagem DL nas tarefas de classificação e predição. Já em Young et al. (2018), os autores comentam que por décadas problemas de Processamento de Linguagem Natural foram abordados com técnicas de aprendizado raso de ML, mas nos últimos anos as redes neurais têm produzido resultados superiores em várias tarefas de NLP.

E.3.1 Modelos e Métodos de *Deep Learning*

Segundo Young et al. (2018), técnicas de aprendizagem profunda utilizam variações de modelos de redes neurais. Os autores identificam e discorrem sobre três modelos fundamentais, que servem de base para as mais variadas formas de implementação: modelo convolucional; modelo recorrente; e, modelo recursivo.

◆ *Convolutional Neural Network (CNN)*

As Redes Neurais Convolucionais são capazes de extrair características salientes da frase de entrada, criando uma representação semântica latente e informativa da frase, para ser utilizada

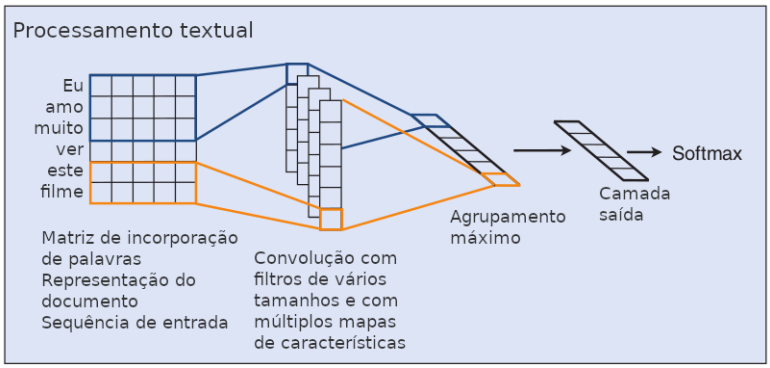
[1] Apesar de imagem não ser o foco deste trabalho, a figura representa bem a diferenciação de modelos ML/DL.

em tarefas posteriores.

A Figura 31 ilustra o modelo básico de CNN, que utiliza técnica conhecida como Abordagem de Sentença. Este modelo normalmente possui centenas de filtros convolucionais, também chamados de *kernels*, de larguras diferentes que "deslizam" sobre a matriz de incorporação de palavras inteira, sendo que cada *kernel* possui função específica para extração de padrões de palavras. Uma camada de convolução é geralmente seguida por uma estratégia de agrupamento máximo (*max pooling*) que fornece uma saída de comprimento fixo. Esta saída de tamanho padronizado reduz a dimensionalidade da saída da rede neural e é normalmente utilizada na tarefa de classificação, pois mantém as palavras características mais salientes da frase.

Já na Abordagem da Janela, que é uma adaptação do modelo básico de CNN, captura-se a estrutura de texto baseado em palavras ao invés de toda a sentença. A marcação de uma palavra depende de suas palavras vizinhas. Considerando o tamanho da janela, selecionam-se as palavras anteriores e posteriores à palavra atual. Esta abordagem é útil principalmente em tarefas de NLP que exigem previsões baseadas em palavras como PoS, NER, e SRL.

Figura 31 – Representação da modelagem de texto por uma CNN.



Fonte: Adaptado de Young et al. (2018).

◆ *Recurrent Neural Network (RNN)*

Devido a natureza sequencial inerente presente na linguagem, este modelo é muito adequado para tarefas de NLP como modelagem de linguagem, tradução automática, reconhecimento de voz e legendagem de imagens. Redes Neurais Recorrentes utilizam a ideia de processar informações sequenciais. O nome recorrente deriva do cálculo repetitivo para cada termo da sequência, bem como da dependência dos cálculos anteriores para realização de cálculos atuais. De certa forma as RNNs têm memória sobre os cálculos anteriores e usam essas informações no processamento atual. O modelo básico de uma RNN é inspirado no modelo proposto inicialmente em Elman (1990).

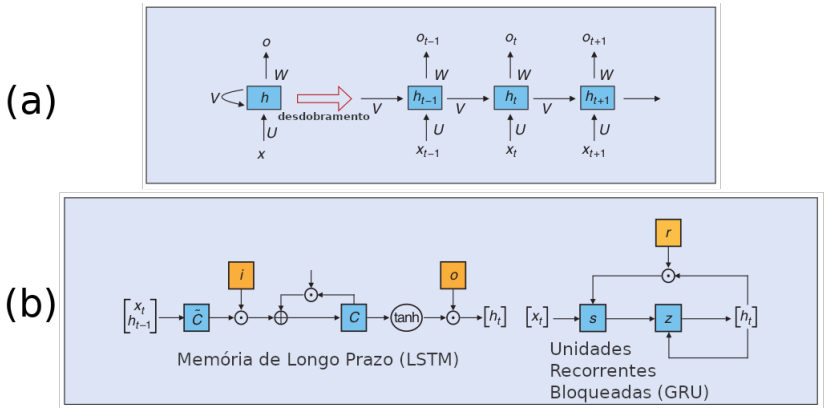
A variante do modelo, denominada Memória de Longo Prazo (LSTM – *Long Short-Term Memory*, possui portas de entrada, saída e uma porta adicional para esquecimento, sendo estas portas responsáveis pelo controle de acesso à memória da camada oculta.

O modelo conhecido por Unidades Recorrentes Bloqueadas (GRU – *Gated Recurrent Units*, também é uma variante com portas de acesso, porém mais simples que o modelo de memória

de longo prazo. Como não tem uma unidade específica para o gerenciamento de memória (o esquecimento), esta variante expõe o conteúdo da camada oculta sem nenhum controle. A GRU possui apenas duas portas (atualização e redefinição) para lidar com o fluxo de informações.

A Figura 32 apresenta: (a) ilustração do modelo básico de RNN baseado na Rede de Elman (1990) e (b) ilustração das variantes do modelo.

Figura 32 – Modelos de Redes Neurais Recorrentes.

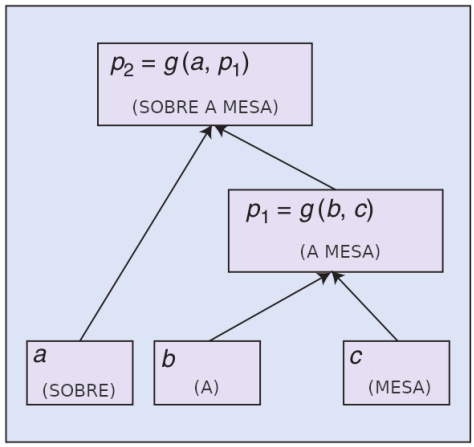


Fonte: Adaptado de Young et al. (2018).

◆ *Recursive Neural Network*

Redes Neurais Recursivas se aproximam da estrutura recursiva natural da linguagem, onde palavras, orações e períodos vão se combinando para formar frases, que consequentemente se combinam na construção de textos. Este é um modelo estruturado em árvore recursiva, de modo que cada nó pai é representado pela composição da representação de seus filhos, esta composição é demonstrada na Figura 33. As características deste modelo o elegem naturalmente para a tarefa de análise sintática.

Figura 33 – Composição do modelo de Rede Neural Recursiva.

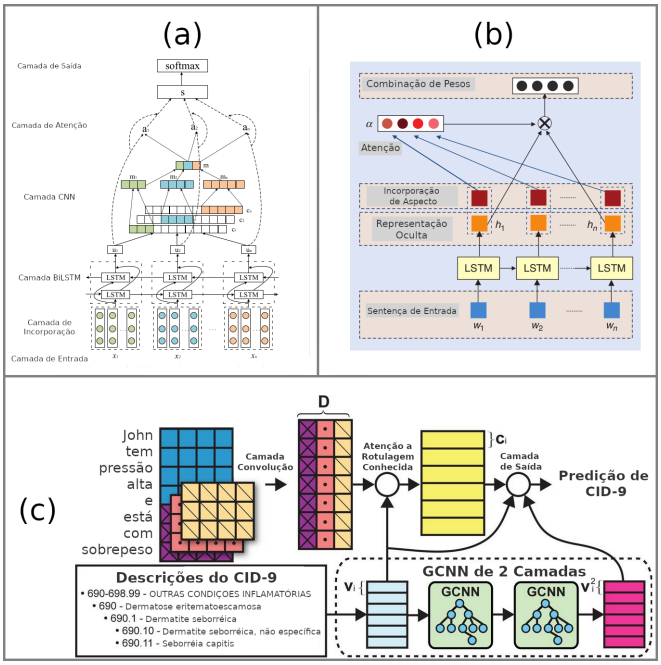


Fonte: Extraído de Young et al. (2018).

◆*Memory-Augmented Networks*

Redes de Memória Aumentada são redes neurais que utilizam Mecanismos de Atenção (*Attention Mechanism*), que proveem informações auxiliares à rede. Estes mecanismos funcionam como *plugins* acoplados na estrutura da rede, com objetivos específicos de acordo com sua implementação. Esta abordagem também é conhecida como *memory-based models* (modelos baseados em memória) (YOUNG et al., 2018).

Figura 34 – Redes neurais com mecanismo de atenção.



Fonte: Adaptado de Zheng e Zheng (2019)(a), Young et al. (2018) (b) e Rios e Kavuluru (2018)(c).

A Figura 34 ilustra três modelos de redes neurais, (a) e (b) são redes recorrentes que utilizam mecanismo de atenção para atribuir pesos diferenciados às características que desejam evidenciar, (c) representa rede neural convolucional que utiliza mecanismo de atenção para destacar palavras presentes num espaço de rótulos estruturados.

E.3.2 Modelagem de documentos

Conforme visto nas Figuras 31 e 34, modelos da abordagem *Deep Learning* utilizam matrizes como entrada de dados. O conceito denominado modelagem de documentos, também chamado de representação de documentos, refere-se a ação de transformar o texto de documentos não estruturados, em estruturas numéricas (e.g., vetores ou matrizes) para serem processadas pelo computador (TAO; CUI; WENJUN, 2018).

◆ *Word Embedding*

O conceito de *Word Embedding (WE)* estabelecido por Mikolov et al. (2013), consiste na representação de palavras em vetores distribuídos no espaço, por isso também é referenciado como representação distribuída. Xiang e Zheng (2018) comentam que a representação de documentos

utilizando WE tem se popular atualmente, por este modelo de representação implicar em melhores resultados nos métodos de classificação.

Apesar de WE poder ser traduzido livremente para "Incorporação de Palavras", observa-se que esta simples tradução é insuficiente para representar a ideia do conceito de *Word Embedding*. Na prática, WE pode ser interpretado como uma extensão do conceito de modelagem/representação de documentos, pois denota a representação de palavras em vetores, de modo que estes vetores capturem relações semânticas entre palavras (pela relação de distância de vetores), ao mesmo tempo que realiza a atribuição de pesos à estas relações, realizando também a redução de dimensionalidade de todo espaço de características, para que seja adequado à entrada do modelo de DL. (TAO; CUI; WENJUN, 2018; XIANG; ZHENG, 2018; YOUNG et al., 2018).

Segundo Maltoudoglou et al. (2022), modelos estáticos (pré-treinados) de *word embeddings*, como o Word2Vec de Mikolov et al. (2013), são criados geralmente por redes neurais rasas de poucas camadas, normalmente redes do tipo convolucional. Ao passo que os modelos dinâmicos, também chamados de contextualizados, são gerados com a utilização de redes neurais profundas e com muitas camadas. Os autores comentam que os modelos dinâmicos conseguem capturar mais informações além das relações semânticas entre palavras.

◆ Modelos generativos

De forma semelhante aos *word embeddings*, existem os modelos sequenciais para representação distribuída de sentenças de texto, como o Seq2Seq de Sutskever, Vinyals e Le (2014). São chamados de modelos generativos por sua capacidade de produção de sentenças completas, sendo estes modelos frequentemente utilizados em tarefas de NLP (e.g., *machine translation*) e geralmente criados com a utilização de redes neurais do tipo recorrente (YOUNG et al., 2018).

Os modelos generativos são compostos por um codificador e um decodificador, onde cabe ao codificador comprimir as informações da sequência de entrada, para que esta seja representada em um vetor de tamanho fixo, chamado de vetor de contexto. No caso de entrada ser do tipo texto, a sequência de entrada refere-se a uma frase ou sentença do texto. O papel do decodificador é ler o vetor de contexto gerado pelo codificador e gerar o mapeamento final da sequência de entrada (MALTOUDOGLOU et al., 2022).

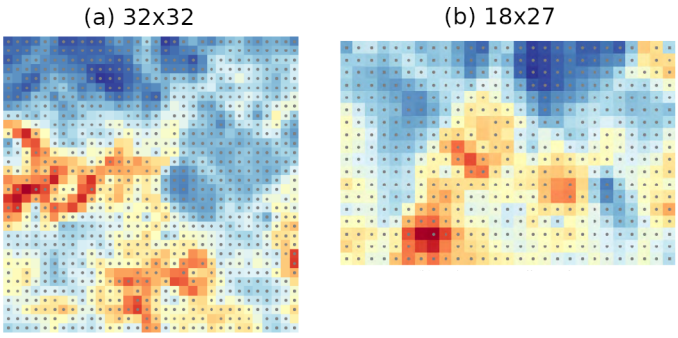
E.4 Modelos e Métodos de Inteligência Artificial

Relembrando a perspectiva sobre Inteligência Artificial (AI) descrita na Seção E.1, que considera ML e DL como parte da AI, mas considerando a categorização realizada por Fisher, Garnsey e Hughes (2016), esta seção apresenta dois modelos que constam tanto nesta revisão sistemática quanto no trabalho referenciado.

No amplo estudo realizado por Fisher, Garnsey e Hughes (2016) — no qual os pesquisadores inspecionaram 262 estudos de análise de texto, relacionados aos domínios de contabilidade, auditoria e finanças — os autores relataram 5 métodos categorizados como AI, nominalmente identificadas por *Expert System (ES)*, *Self-Organizing Maps (SOM)*, *Fuzzy Neural Network (FNN)*, *Genetic Algorithm (GA)* e *Intelligent Agents (IA)*.

- *Self-Organizing Map (SOM)* — O Mapa Auto-Organizável é um modelo que se destaca pela característica de realizar agrupamentos de forma automática com base na semelhança vetorial das instâncias. Neste modelo, vetores de altas dimensões são naturalmente mapeados para vetores bidimensionais, formando clusters identificáveis. Com relação a tarefa de classificação, a rotulagem das amostras também é realizada automaticamente com a utilização dos rótulos presentes nos clusters identificados (GIANNOPOULOU; MITROU, 2018).

Figura 35 – Modelos de *grids* para SOM.



Fonte: Adaptado de Giannopoulou e Mitrou (2018).

A Figura 35 apresenta dois tipos de *grids* — (a) quadrado, (b) não-quadrado — que são modelos estruturais nos quais o método SOM organiza automaticamente as instâncias analisadas. Observa-se na figura a identificação de clusters sendo evidenciada pelo mapa de calor. A presença do ponto cinza no espaço dimensional do mapa, revela a atribuição de rótulo aquele espaço, consequentemente as amostras que se enquadram no espaço recebem a mesma rotulação (GIANNOPOULOU; MITROU, 2018).

- *Genetic Algorithm (GA)* — Algoritmos Genéticos são baseados nos princípios da seleção natural e do melhoramento genético, características que tornam este método naturalmente adequado para problemas de otimização. A implementação de GA baseia-se na codificação de potenciais soluções para um problema, em forma de cromossomos. Onde cada solução (cromossomo) é um indivíduo nesta abordagem, de modo que o conjunto deles compõe a população de possíveis soluções. A partir desta modelagem, realizam-se operações de seleção, codificação, cruzamento e mutação na população, com objetivo de cobrir todas soluções possíveis e suas descendências (MUSTAFI; MUSTAFI; SAHOO, 2020).

Tanto no estudo de Giannopoulou e Mitrou (2018) utilizando SOM, quanto no estudo de Mustafi, Mustafi e Sahoo (2020) com GA, os modelos utilizados se enquadram no paradigma de aprendizagem não supervisionada. Ambos estudos aplicam suas técnicas prioritariamente para satisfazer a tarefa de clusterização, sendo o segundo estudo um caso de dedicação exclusiva para tal tarefa. Desta forma, somente o trabalho de Giannopoulou e Mitrou (2018) será considerado na análise de técnicas e métodos para tarefa de classificação de texto.

E.5 Considerações finais sobre a Inteligência Artificial

Este capítulo discorreu sobre alguns dos conceitos e abordagens de AI, descrevendo fundamentos intrinsecamente relacionados com a implementação da Classificação de Texto e do Processamento de Linguagem Natural. Os conceitos, técnicas, modelos e métodos apresentados aqui, foram majoritariamente referenciados por estudos incluídos na revisão sistemática.