UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Gabriel Rücker

Qualidade de dados como requisito na comparação de algoritmos de classificação de conteúdo textual

São José do Rio Preto
2022

Gabriel Rücker

Qualidade de dados como requisito na comparação de algoritmos de classificação de conteúdo textual

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Orientadora: Profa. Dra. Adriana Barbosa Santos

São José do Rio Preto
2022

R911q

Rücker, Gabriel

Qualidade de dados como requisito na comparação de algoritmos de classificação de conteúdo textual / Gabriel Rücker. -- São José do Rio Preto, 2022

76 f. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientadora: Adriana Barbosa Santos

Coorientador: Álvaro Magri Nogueira da Cruz

1. Ciência da computação. 2. Mineração de dados (Computação). 3. Inteligência artificial. 4. Big data. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Gabriel Rücker

Qualidade de dados como requisito na comparação de algoritmos de classificação de conteúdo textual

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Comissão Examinadora

Profa. Dra. Adriana Barbosa Santos
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof. Dr. Adriano Mauro Cansian
UNESP – Câmpus de São José do Rio Preto

Prof. Dr. Leandro Alves Neves
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
04 de janeiro de 2022

Agradecimentos

Gostaria de agradecer aos meus pais, Ricardo e Márcia, que sempre me apoiaram para que eu seguisse em frente e forneceram todas as oportunidades para que me desenvolvesse e chegasse até aqui. Agradeço também à minha companheira, Gabriela, que sempre esteve ao meu lado em todos os anos que estamos juntos, ajudando a superar as diferentes fases e dificuldades que se apresentaram no decorrer dessa trajetória.

Sou grato a todos os professores com os quais tive contato e oportunidade de aprender. Apesar da experiência do ensino remoto durante dois anos da graduação, o conhecimento transmitido pelos docentes nunca deixou de ser excelente, o que comprova a qualidade da instituição. Em especial, agradeço a Prof.ª Dr.ª Adriana Barbosa Santos, a qual orientou-me desde o início e contribuiu de forma significativa para que este trabalho fosse elaborado da melhor forma possível. Agradeço também ao Prof. Me. Álvaro Magri Nogueira da Cruz, o qual auxiliou na indicação e elucidação de diversos conceitos que foram incorporados neste trabalho.

Por fim, agradeço a todos aqueles com os quais compartilhei experiências e vivências, seja de maneira presencial ou virtual, nas diversas atividades acadêmicas e extra-curriculares da universidade, pois contribuíram na minha formação como indivíduo e estarão presentes de alguma forma nessa nova jornada.

Resumo

A disponibilidade de dados é cada vez maior no mundo atual. Para a realização de inferências que auxiliem nas tomadas de decisão, técnicas de mineração de dados como os algoritmos de classificação são utilizadas. Além disso, a qualidade dos dados é um fator preponderante, pois influencia diretamente nos resultados que embasam o processo decisório, especialmente no ambiente corporativo. Este trabalho objetiva a comparação de algoritmos que priorizem a avaliação de dimensões de qualidade de dados no processo de classificação de conteúdo textual a partir de artigos científicos que compõem o portfólio de informações exibidas em uma plataforma de comunicação científica. Para isso, utilizaram-se diferentes técnicas de validação para aferir o desempenho dos algoritmos, bem como métricas específicas para avaliação das dimensões de qualidade sob diferentes condições experimentais, visando avaliar a influência da qualidade dos dados no processo de classificação.

Palavras-chave: Qualidade de dados. Mineração de dados. Algoritmos de classificação. Dimensões de qualidade de dados. Certificação de qualidade dos dados.

Abstract

Data availability is growing in the real world. In order to be able to make inferences that help in decision making, data mining techniques such as the classification algorithms are used. Besides that, data quality is a relevant factor to consider, because it directly impacts results of the decision-making process, especially in the corporative environment. This work aimed to compare algorithms that prioritize data quality dimensions evaluation in the process of classifying textual content from scientific papers for a scientific communication platform. In order to do that, different validation techniques were used to measure the algorithms performance, as well as specific metrics to evaluate the quality dimensions under varied experimental conditions, aiming to evaluate data quality impact on the outcomes of a classification.

Keywords: Data quality. Data mining. Classification algorithms. Data quality dimensions. Data quality certification.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Considerações Iniciais

Estudos envolvendo a mineração de dados têm crescido nos últimos anos, pois a quantidade de informações disponíveis tem aumentado de forma significativa a partir de dados não estruturados como os textos, presentes em documentos, artigos de notícias, correios eletrônicos, repositórios governamentais e de blogs, além de tantos outros (BAHARUDIN et al., 2010). Portanto, tornou-se importante a tarefa de extração de padrões e tendências desconhecidas a partir de grandes bases de dados por meio da mineração de dados (BRINDHA; PRABHA; SUKUMARAN, 2016).

Diversas metodologias são utilizadas em conjunto para se extrair conhecimento a partir dos dados. Caso a classificação seja de dados textuais, combinam-se as técnicas de Processamento de Linguagem Natural (PLN), mineração de dados e aprendizado de máquina para a obtenção de resultados (BAHARUDIN et al., 2010). Porém, o processo de classificação não está limitado apenas a esse tipo de dado, podendo ainda ser utilizado para a descoberta de conhecimento a partir de conteúdos multimídia, por exemplo. Segundo Vaughan (1993) apud (GE; PERSIA, 2017), "Multimídia é qualquer combinação de texto, arte gráfica, som, animação, e vídeo proporcionada por um computador".

Outro problema decorrente da grande disponibilidade de dados e sua heterogeneidade,

característicos da era do *Big Data*, é a necessidade, por parte de pesquisadores e tomadores de decisões, em melhorar a qualidade dos dados manipulados, pois os dados são a principal fonte de informação sobre os indicadores estabelecidos para compreensão das necessidades dos consumidores, qualidade dos serviços oferecidos e na mitigação de possíveis riscos decorrentes das deficiências de qualidade na base de dados, típicas do ambiente do ambiente corporativo (CAI; ZHU, 2015).

Dessa maneira, é pertinente abordar o fato de que, ao se projetar um sistema que dependa diretamente dos dados obtidos para atingir um desempenho satisfatório, como aqueles baseados em algoritmos de recomendação ou classificação, por exemplo, deve-se priorizar a qualidade dos dados sobre o qual o sistema se apoia (HEINRICH et al., 2019). Segundo Karr, Sanil e Banks (2003), qualidade de dados é a capacidade de utilizar as informações de maneira efetiva, econômica e rápida para informar e avaliar decisões. Para mensurar e avaliar a qualidade dos dados, as dimensões de qualidade dos dados representam uma opção viável para tal (MCGILVRAY, 2008).

A problemática que considera os pontos acima expostos, isto é, a necessidade de se garantir a qualidade dos dados de sistemas que fazem a mineração de dados, em especial os algoritmos de classificação, norteia a proposição deste trabalho, cujos objetivos são apresentados a seguir.

1.2 Objetivo

Este projeto objetiva a comparação de algoritmos, priorizando a avaliação de dimensões de qualidade de dados no processo de classificação de conteúdo textual proveniente de artigos científicos que compõem o portfólio de informações de uma plataforma de comunicação científica. Mais especificamente, este trabalho visa:

- Realizar a seleção de algoritmos de classificação a partir de uma base de dados utilizada para testes;

- Avaliar o efeito da inserção de ruídos sobre a acurácia dos algoritmos de classificação, analisando diferentes dimensões de qualidade de dados.

1.3 Motivação

Recentemente, considerando a grande disponibilidade de dados e o avanço na era do *Big Data*, a qualidade dos dados tornou-se um fator importante para qualquer negócio disposto a utilizar as informações armazenadas em seus repositórios. Para se ter ideia, em 2016 a estimativa para o tamanho do mercado de *Big Data* era de 136 bilhões de dólares. Apesar disso, no mesmo ano, o custo estimado de perda anual, apenas nos EUA, decorrente da baixa qualidade de dados, foi de 3,1 trilhões de dólares (REDMAN, 2016).

Dessa maneira, Laney (2017) apud (GUALO et al., 2021) e Redman (2013) encontraram associações entre a baixa qualidade de dados e diversos problemas organizacionais, apontando que:

- A qualidade de dados pode afetar a produtividade no trabalho em até 20%;
- A baixa qualidade dos dados é a razão primária pela qual 40% de todos os negócios falham em atingir metas de lucro;
- A qualidade dos dados pode ser um fator limitante para melhoria da qualidade dos processos quando estes se tornam automatizados.

Por outro lado, ainda no contexto do *Big Data*, existe a necessidade de se extrair padrões e tendências a partir de grandes conjuntos de dados que auxiliem na tomada de decisão. Nesse sentido, a mineração de dados e suas técnicas, como a classificação, permitem que as informações obtidas sejam utilizadas em diversas áreas, tais como análise de mercado, controle de produção, detecção de fraudes, análise de dados científicos e outros (BRINDHA; PRABHA; SUKUMARAN, 2016).

Blake e Mangiameli (2011) uniram os conceitos envolvendo qualidade de dados com a mineração de dados e concluíram que a complexidade dos problemas de classificação, a qual possui interação com aspectos envolvendo a qualidade de dados, interferem diretamente nos resultados de classificação.

Logo, é preciso avaliar a qualidade dos dados no contexto de como as mudanças desses dados impactam nos desfechos de classificação, pois, desta maneira, é possível adotar medidas para mitigar o impacto da baixa qualidade no processo de classificação e, consequentemente, para a aplicação na solução de casos reais.

Capítulo 2

Fundamentação Teórica

Com o crescimento da Internet e, consequentemente, da disponibilidade de dados não estruturados e semi-estruturados, houve uma crescente demanda por algoritmos que classificassem automaticamente as informações como forma de organizá-las (BAHARUDIN et al., 2010).

Outro efeito notável desse aumento do volume de dados é a preocupação crescente das organizações em priorizar a qualidade dos dados, pois há mais chances de interpretar com transparência os indicadores que neles se baseiam e que dão suporte para a tomada de decisão por parte dos gestores (FLECKENSTEIN; FELLOWS, 2018).

Nesse sentido, Blake e Mangiameli (2011) procuraram realizar uma associação entre os problemas de classificação e a qualidade dos dados. Os autores concluíram que a complexidade do problema é fator determinante para a qualidade de dados, de tal forma que a estrutura de um problema de classificação é capaz de ampliar os efeitos negativos em um conjunto de dados de baixa qualidade.

Logo, é necessário que as dimensões de qualidade dos dados sejam priorizadas na elaboração de um algoritmo de classificação, visto que as mesmas podem interferir no resultado final, podendo acarretar em prejuízos difíceis de mensurar aos negócios.

Visando fornecer o embasamento necessário para compreensão do tema, neste capítulo são apresentados conceitos envolvendo algoritmos de classificação, especificamente aqueles

que lidam com dados textuais, as dimensões de qualidade dos dados e a série de certificações ISO/IEC 25000, que oferece à indústria critérios para a garantia da qualidade dos dados armazenados em seus repositórios. Por fim, é realizada uma comparação entre os algoritmos mais utilizados para os problemas de classificação.

2.1 Algoritmos de classificação

Os algoritmos de classificação vêm sendo utilizados em uma variedade de áreas, como banco de dados, mineração de dados e extração de informações. Tipicamente, utiliza-se um conjunto de registros de treino para criar um modelo de classificação que associa as características de um determinado registro a uma classe específica, sendo esta denotada por um rótulo que será fornecido àquele registro após a classificação (AGGARWAL; ZHAI, 2012). O modelo construído é então utilizado para classificar novos registros submetidos ao sistema, por meio da aplicação de um algoritmo de classificação.

Uma tarefa de classificação pode utilizar tipos de dados variados, como: dados comerciais, textos, DNAs e imagens. Mais ainda, as classificações podem ser binárias, multi-classe ou multi-rotuladas (HOSSIN; SULAIMAN, 2015). Além disso, uma diversidade de algoritmos podem ser usados para classificar um determinado objeto. Zhongguo et al. (2017) proporam um algoritmo de recomendação de classificadores, ressaltando que a escolha dependerá das características do conjunto de dados com o qual se está trabalhando. Para isso, sugeriram o seguinte método:

1. Predizer a variação de desempenho para diferentes algoritmos sobre um mesmo conjunto de dados;
2. Predizer o melhor algoritmo obtido;
3. Predizer os parâmetros ideais para o algoritmo.

Dessa maneira, pode-se dizer que as tarefas de classificação envolvem tanto o tipo de dado que será utilizado como parâmetro, como também as características do conjunto de dados.

Logo, a seleção de um classificador apropriado deve ser a partir da avaliação de ambos os critérios.

2.1.1 Técnicas de classificação de textos

A mineração de textos é a detecção, a partir de um computador, da informação mais próxima e desconhecida até então para a extração automática de padrões ou conhecimentos a partir de dados não estruturados presentes em uma variedade de fontes (BRINDHA; PRABHA; SUKUMARAN, 2016).

De acordo com Vasa (2016), a classificação de textos pode ser dividida em quatro fases, como mostra a Figura 2.1:

- Pré-processamento de texto;
- Extração de características;
- Classificador de treino;
- Modelo de classificação.

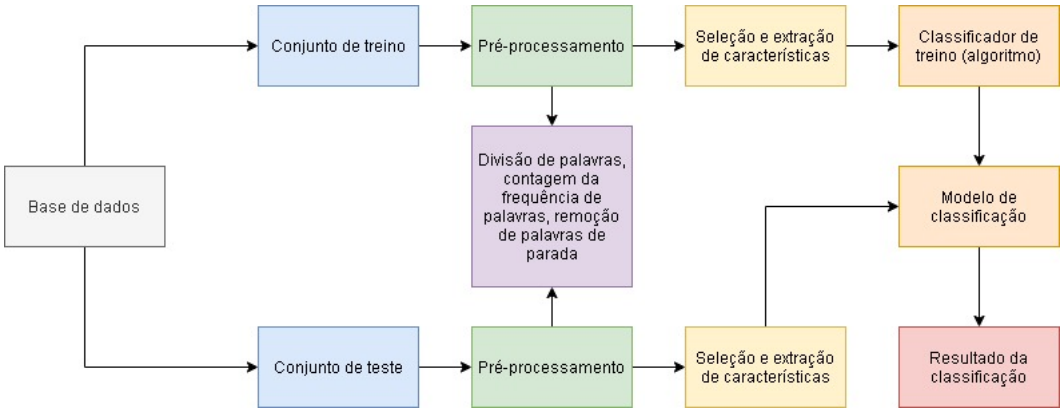
O conjunto de dados é inicialmente dividido em um conjunto de treino e outro de teste, sendo que cada um segue a mesma sequência de passos até a classificação.

Em seguida, tem-se o pré-processamento de dados, o qual constitui um passo mandatório no processo de mineração de dados, pois é responsável por converter dados inúteis em algo que permita a extração de características. Caso isso não ocorra, é possível que o algoritmo apresente erros durante a execução ou, caso funcione, apresente resultados incoerentes com o que foi proposto inicialmente (GARCíA; LUENGO; HERRERA, 2014). Conforme mostra a Figura 2.1, este processo pode ser realizado por meio da divisão das palavras de um texto, contagem da frequência de palavras e remoção de palavras de parada, as quais são geralmente irrelevantes durante a classificação.

Uma limitação inerente à classificação de textos é a alta dimensionalidade do espaço de características, devido ao elevado número dessas. Consequentemente, a complexidade dos algoritmos tende a aumentar e a acurácia diminuir, pois existe a presença de termos irrelevantes ou redundantes nos dados. Uma maneira de se lidar com isso é utilizar a seleção e extração de características (SHAH; PATEL, 2016).

Por fim, seleciona-se um classificador para o conjunto de treino, ou seja, o algoritmo que será utilizado para a classificação. Feito isso, o conjunto de teste é incluído para a validação e criação de um modelo de classificação, o qual será utilizado para as classificações efetivamente.

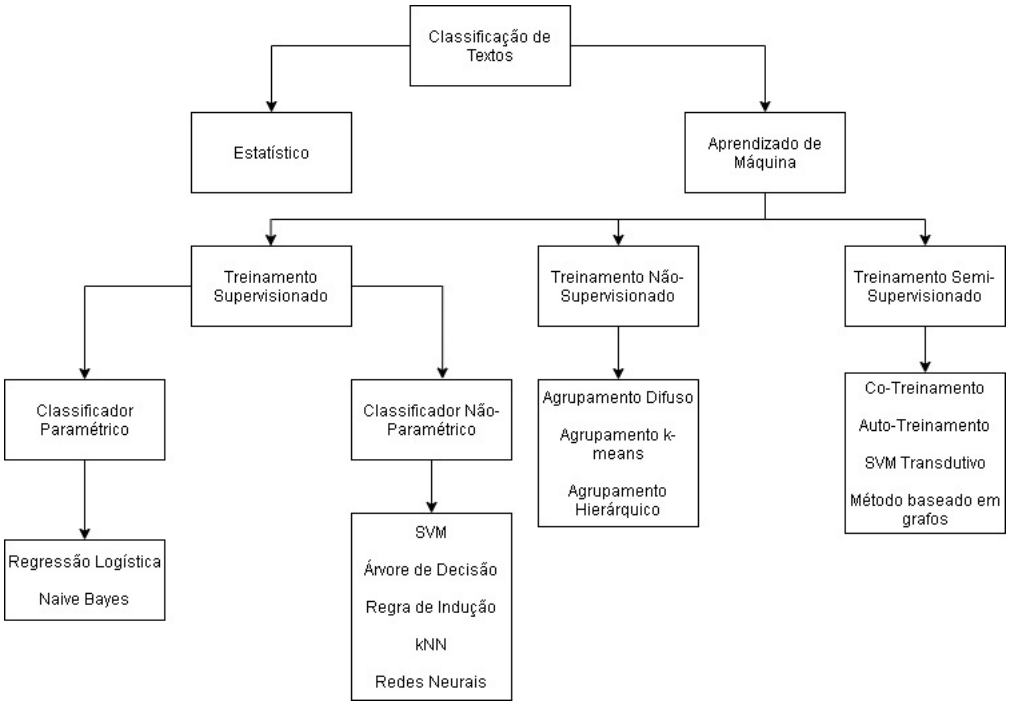
Figura 2.1: Processo de construção de um classificador de textos.



Fonte: Adaptado de Vasa (2016).

Quanto à disponibilidade de classificadores, Thangaraj e Sivakami (2018) identificaram e descreveram cada uma das técnicas apresentadas na Figura 2.2, destacando as principais diferenças e similaridades entre cada uma das abordagens, bem como os algoritmos que as descrevem.

Figura 2.2: Técnicas de classificação de textos.



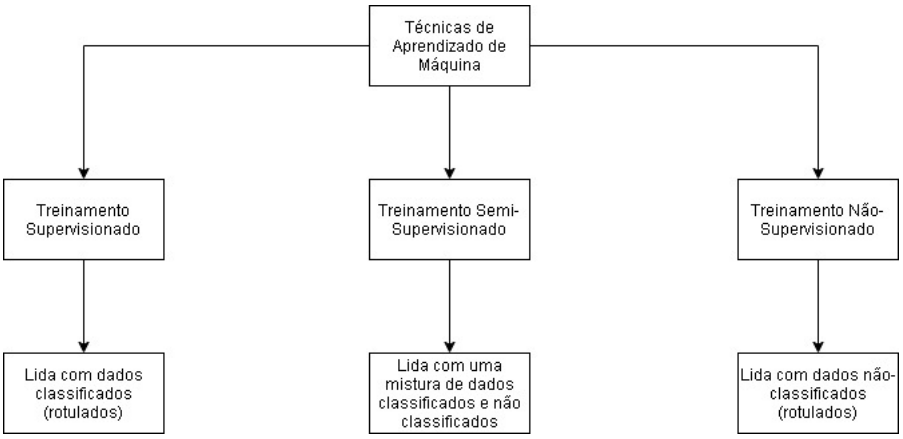
Fonte: Adaptado de Thangaraj e Sivakami (2018).

Sendo assim, a classificação de textos pode ser dividida de duas formas: utilizando abordagens estatísticas e algoritmos de aprendizado de máquina.

Quanto aos métodos estatísticos, estes se baseiam na formalização das relações entre variáveis na forma de equações matemáticas (SRIVASTAVA, 2015). Como exemplo, Seara Vieira, Borrajo e Iglesias (2016) propõem um método matemático para a redução da dimensionalidade, importante durante a etapa de pré-processamento.

Os algoritmos de aprendizado de máquina, por sua vez, subdividem-se em três categorias: treinamento supervisionado, treinamento semi-supervisionado e treinamento não-supervisionado. A Figura 2.3 apresenta a diferença entre cada um na forma de tratamento dos dados.

Figura 2.3: Diferentes tipos de aprendizado de máquina.



Fonte: Adaptado de Mohammed, Khan e Bashier (2016).

Dessa forma, o treinamento supervisionado utiliza dados rotulados, a partir de máquinas ou humanos, para a classificação. Já no treinamento não-supervisionado não há supervisores ou dados de treinamento, apenas dados não rotulados. Por fim, o treinamento semi-supervisionado utiliza ambas as técnicas para a classificação.

Verifica-se que dentre os procedimentos estatísticos há uma sub-divisão dos classificadores no treinamento supervisionado, podendo estes serem paramétricos ou não-paramétricos. Os testes paramétricos são baseados em suposições feitas a respeito dos parâmetros da distribuição da população, inferindo sobre eles a partir de dados amostrais. Já os testes não-paramétricos não assumem uma forma ou parâmetros para a distribuição da população (HOSKIN, 2012).

Como cada algoritmo de classificação de texto possuirá vantagens e desvantagens dentro do contexto de uma aplicação específica, é preciso conhecer o objetivo da classificação para escolher um algoritmo que melhor atenda às necessidades. A Tabela 2.1 descreve alguns dos classificadores mais utilizados para o treinamento supervisionado:

Tabela 2.1: Vantagens, desvantagens e aplicações de algumas técnicas de classificação de textos.

Método	Vantagens	Desvantagens	Aplicações
Regressão Logística	Estimação simples dos parâmetros, funciona bem para predições categóricas.	Requer amostras grandes, inútil para problemas não-lineares, suscetível ao excesso de confiança.	Predição do custo de software, projeção financeira, mineiração de dados de crimes.
Naive Bayes	Classificador rápido, converge mais rápido que a regressão logística, necessita de menos treinamento, é aplicável para problemas binários e multi-classe	Interações entre as características não acontece. Probabilidades calculadas não são matematicamente exatas, mas relativas.	Marcação de e-mails, classificação de artigos baseados no conteúdo, análise de sentimentos.
SVM	Parâmetros de regularização previnem o *over-fitting*, a engenharia de núcleo ajuda na incorporação de conhecimento especializado.	Seleção do melhor núcleo e o tempo despendido treinando e testando.	Bases de dados biológicas, categorização de hipertexto, entre outras.
k-NN	Implementação simples, bom desempenho em problemas multi-classe e flexível na seleção de características.	Procura dos vizinhos mais próximos e estimativa do melhor valor para k.	Sistemas de recomendação.
Redes Neurais Artificiais	Mais fácil de utilizar, aproxima qualquer tipo de função e chega próximo ao cérebro humano.	Requer um grande conjunto de teste e treinamento, muitas operações não transparentes e dificuldade de aumentar a acurácia.	Projeção de vendas, validação de dados, gestão de riscos e marketing direcionado.

Fonte: Adaptado de Thangaraj e Sivakami (2018).

2.2 Dimensões de qualidade dos dados

Segundo McGilvray (2008), uma dimensão de qualidade de dados é uma característica ou parte da informação utilizada para classificar informação e requisitos dos dados.

Existem muitas definições acerca das diferentes dimensões de qualidade dos dados, e esta discrepância na definição de cada uma delas ocorre devido à natureza contextual de qualidade. Além disso, não há um consenso geral de quais dimensões constituem o conjunto mais importante, tampouco uma definição exata para cada uma delas (BATINI et al., 2009).

Dito isso, as dimensões podem ainda ser agrupadas em hiperdimensões. Karr, Sanil e Banks (2003) identificaram três delas:

- Processo: geração, montagem, descrição e manutenção dos dados. Dimensões que de forma geral são avaliadas subjetivamente e qualitativamente, como: confiabilidade, metadados, segurança e confidencialidade.

- Dados: dimensões associadas com os dados em si, medidas de forma quantitativa, como: acurácia, completude, consistência e validade.
- Usuário: dimensões associadas ao usuário e usabilidade de forma geral. São medidas qualitativas, como: acessibilidade, integrabilidade, interpretabilidade, retificabilidade, relevância e temporalidade.

Por esse enfoque, o estudo das dimensões de qualidade são abordadas dentro do contexto das hiperdimensões, de forma a contextualizar melhor a qualidade dos dados, visto que este é um assunto abrangente e necessita de um escopo bem definido. Algumas das principais dimensões são detalhadas a seguir.

2.2.1 Acurácia

A acurácia de um dado pode ser avaliada verificando se os valores armazenados em uma base de dados correspondem aos respectivos valores reais (BALLOU; PAZER, 1985). Além dessa definição, outras podem ser dadas para esta dimensão, conforme mostra a Tabela 2.2. Por ser uma dimensão quantitativa, pode ser obtida por meio da razão entre os dados corretos e os dados válidos da base de dados, conforme a equação a seguir:

$$\text{Acurácia} = \frac{D_c}{D_v} \tag{2.1}$$

Na Eq. (2.1), D_c representa os dados corretos, enquanto D_v os dados válidos de uma determinada base de dados.

Batini et al. (2009) classificaram a acurácia em duas categorias: sintática e semântica. As metodologias que utilizam a qualidade de dados costumam utilizar apenas a forma sintática de avaliação da acurácia, sendo esta definida como a aproximação de um determinado valor v, aos elementos do domínio D.

Tabela 2.2: Definições de acurácia a partir de diferentes autores.

Referência	Definição
Wand and Wang (1996)	Habilidade de um sistema de informação para representar todo estado significativo de um sistema real
Wang and Wand (1996)	Extensão em que os dados possuem largura, profundidade e escopo suficientes para a tarefa em questão
Redman (1996)	Grau em que valores são incluídos em uma coleção de dados
Jarke et al. (1995)	Porcentagem de informações reais que entraram em fontes de dados e/ou armazém de dados
Bovee et al. (2001)	Informação possuindo todas as partes necessárias da descrição de uma entidade
Naummann (2002)	Razão entre o número de valores não-nulos em uma fonte e o tamanho da relação universal
Liu and Chi (2002)	Todos os valores que devem ser coletados conforme uma teoria da coleção.

Fonte: Adaptado de Batini et al. (2009).

Dessa forma, a acurácia pode ser difícil de ser medida dependendo do contexto em que está inserida. Ferreira (2020) evidencia a diferença de se avaliar esta dimensão por meio de dois exemplos: comparação dos dados correspondentes ao estoque físico de mercadorias e aqueles observados na base de dados, e a descrição de um determinado produto em um site de comércio eletrônico. Enquanto naquele o cálculo da acurácia é mais fácil por ser exato, neste pode haver dificuldade em determinar se a informação é equivalente à realidade.

2.2.2 Completude

Completude é o grau em que uma determina coleção de dados inclui as informações que descrevem o conjunto de objetos reais correspondentes (BATINI et al., 2009). Apesar de parecer uma dimensão não-ambígua, pois um dado pode possuir valores ausentes ou não, problemas podem surgir ao confundir dados não presentes com a legitimidade destes (KARR; SANIL; BANKS, 2003). A completude pode ser expressa pela seguinte equação:

$$\text{Completude} = \frac{N}{T} \tag{2.2}$$

Na Eq. (2.2), os valores não nulos são expressos por N e T representa o totalidade de dados presente na base.

Brassel et al. (1995) chegaram à conclusão que existem dois tipos de completude: a de dados e a de modelo. A completude de dados é um erro de omissão que pode ser mensurado entre a base de dados e sua especificação, podendo ser utilizada independentemente de uma aplicação. Já a completude de modelo depende da aplicação em questão. Costuma estar associada à conformidade da especificação da base de dados e a descrição concreta que é necessária para uma aplicação específica de banco de dados.

Um exemplo desta dimensão está na exigência, por parte de um site que exige um cadastro, por exemplo, que certos campos sejam preenchidos para que o usuário seja efetivado pelo sistema (FERREIRA, 2020).

2.2.3 Consistência

A consistência é baseada na verificação da violação de regras semânticas que regem uma base de dados. Um tipo de regra semântica utilizada pela teoria relacional é a de restrições de integridade. Esta, por sua vez, pode ser subdividida em duas: as restrições inter-relações e as restrições intra-relações. Enquanto esta assume um intervalo de valores válidos para um atributo do domínio, aquela envolve atributos de múltiplas relações (BATINI et al., 2009).

Como forma de tentar avaliar esta dimensão de maneira quantitativa, Alpar e Winkelsträter (2014) proporam uma métrica para a consistência por meio da seguinte tupla:

$$\text{Consistência}(t) := \sum_{r \in R} \begin{cases} w^{+}(r), & \text{se t obedece r} \\ w^{-}(r), & \text{se t viola r} \\ w^{0}(r), & \text{se r não se aplica} \end{cases} . \tag{2.3}$$

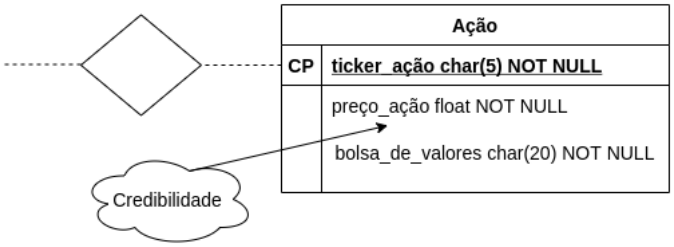
Na Eq. (2.3), R é um conjunto de regras de associação, $w^{+}(r)$ significa que determinada regra de associação foi obedecida, aumentando uma pontuação para a consistência, e $w^{-}(r)$ que a mesma foi violada, subtraindo pontos. Por fim, $w^{0}(r)$ significa que a regra de associação não é aplicável, fazendo com que a pontuação permaneça inalterada. Desta maneira, tuplas com mais pontos também são, por consequência, mais consistentes que as outras.

Assim, apesar da avaliação quantitativa ser relativamente simples, a consistência pode ser difícil de ser mensurada, pois nem sempre envolve diretamente um único atributo. Uma das maneiras utilizadas para a correção de inconsistências nas bases de dados é feita procurando-se por atributos que possibilitem a correção dos dados. Como exemplo, uma pessoa pode ter o CEP e a cidade corretos em uma base de dados, porém se o estado não condiz com essas informações, é possível corrigir apenas o estado (FERREIRA, 2020).

2.2.4 Credibilidade

A credibilidade pode ser medida pela avaliação do quanto se acredita que uma informação seja verdadeira e confiável por parte dos usuários dentro de um determinado contexto de uso (GUALO et al., 2021). Como exemplo desta dimensão, considere uma base de dados contendo informações a respeito dos preços das ações da bolsa de valores em um determinado dia do ano (WANG; KON; MADNICK, 1993). Um usuário que pretende fazer uso daquelas informações para realizar operações necessita que as informações presentes possuam credibilidade, ou seja, reflitam o preço real naquele instante e sejam confiáveis para que o usuário possa utilizá-las. Para ilustrar melhor esta situação, a Figura 2.4 mostra a entidade ação e seus atributos, bem como a necessidade da credibilidade na definição do preço de uma determinada ação.

Figura 2.4: Descrição da entidade ação e relação com a credibilidade.



Fonte: Adaptado de Wang, Kon e Madnick (1993).

Redman (2013) examinou a problemática que envolve a credibilidade e cita estudos sobre trabalhadores qualificados que dependem até 50% de seu tempo de trabalho na identificação e correção de erros tentando confirmar as fontes para os dados que eles não confiam completamente. Dessa forma, o autor conclui que a solução para isso não é melhorar as tecnologias existentes. É preciso mudar o foco da responsabilidade: em vez de enfocar nos profissionais de tecnologia, os quais não estão ligados diretamente ao processo de criação dos dados, deveria enfocar nos administradores do negócio, os quais, em tese, são mais aptos a obter os dados de forma adequada.

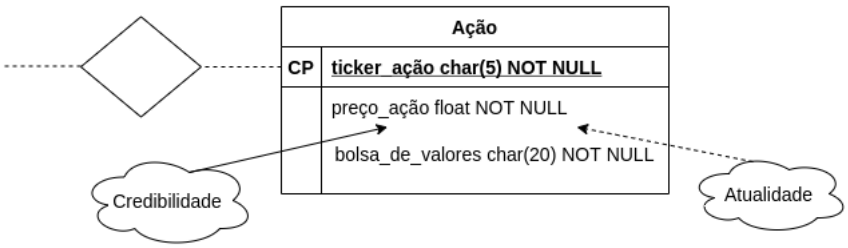
2.2.5 Atualidade

A atualidade pode ser avaliada por meio da verificação dos atributos de um dado e seu respectivo tempo de existência dentro de um determinado contexto de uso (GUALO et al., 2021). Para entender a importância desta dimensão em uma situação onde há a presença de um armazém de dados, Li, Peng e Kennedy (2010) proporam um método de classificar a impureza dos dados do ponto de vista das dimensões de qualidade, chegando à conclusão que as dimensões mais importantes de se avaliar são a acurácia e a atualidade, pois no caso de um sistema bancário online, por exemplo, os dados precisam ter uma determinada exatidão e estarem atualizados o suficiente para que as informações disponibilizadas para os usuários estejam corretas.

Voltando para o exemplo de um usuário operando na bolsa de valores, a atualidade dos

dados seria necessária neste caso, pois os preços das ações devem estar suficientemente atualizados para que as operações aconteçam de forma simultânea, ou bem próximo disso. A Figura 2.5 ilustra o mesmo caso descrito no sub-tópico anterior, agora com a dimensão atualidade presente.

Figura 2.5: Descrição da entidade ação com as dimensões de qualidade atualidade e credibilidade.



Fonte: Adaptado de Wang, Kon e Madnick (1993).

2.3 Desafios na avaliação da qualidade dos dados

Em Fleckenstein e Fellows (2018) são apresentados alguns desafios envolvendo a qualidade dos dados, sobretudo no ambiente de negócios, como consequência da falta de uma abordagem adequada para administração dos dados. São eles:

- Controles inadequados no ponto de origem;
- Volume, variedade e velocidade dos dados;
- Complexidade do ambiente operacional;
- Proliferação e duplicação excessiva dos dados;
- Metadados pobres, definições incertas e múltiplas interpretações.

O problema de controles inadequados no ponto de origem pode ocorrer com dados obtidos externamente ou internamente. No caso deste, os erros podem se originar manualmente no

momento de entrada dos dados, bem como devido à ineficiência na definição e aplicação das regras de negócio em sistemas transacionais antes de originar conteúdo. No caso daquele, pode acontecer por meio de controles inadequados de aquisição, Acordos de Nível de Serviço (ANS) mal definidos e ausência de um processo de governança pelo qual os dados são trazidos para a organização (FLECKENSTEIN; FELLOWS, 2018).

Devido ao avanço na era do *Big Data*, tornou-se necessário avaliar a qualidade dos dados levando em consideração suas características. Sendo assim, volume, variedade e velocidade representam o volume de dados crescente, a variedade de informações na forma de dados estruturados e não-estruturados, bem como a necessidade de processamentos de alto desempenho para lidar com uma grande massa de dados, respectivamente (CAI; ZHU, 2015).

A complexidade do ambiente operacional se deve, sobretudo, ao avanço da tecnologia relacionada aos sistemas distribuídos, como, por exemplo, a computação em nuvem. Dessa maneira, os desafios se mostram presentes ao buscar coerência entre sistemas dispersos, estruturas, sistemas de originação e volumes limitados (FLECKENSTEIN; FELLOWS, 2018).

Como os dados, com o passar do tempo, foram custando cada vez menos para serem armazenados, devido ao avanço das tecnologias de armazenamento e o aumento na quantidade de dados disponíveis, as empresas puderam adquirir mais conjuntos de dados de forma ampla e fácil, contribuindo para sua proliferação. Além disso, problemas podem surgir devido à tendência coletiva em duplicar dados, gerando ainda mais dificuldade em administrá-los (FLECKENSTEIN; FELLOWS, 2018).

Por fim, os metadados, os quais representam as informações a respeito dos dados de uma organização, são comumente extraídos de forma técnica sem prover aplicabilidade nos negócios. Tais problemas podem ocorrer pela falta de prioridades, por parte da organização, em desenvolver e documentar os dados criados, adquiridos e compartilhados, causando confusão quanto ao significado dos dados e sua utilização (FLECKENSTEIN; FELLOWS, 2018).

2.4 Certificação de qualidade dos dados ISO/IEC 25000

Segundo García, Luengo e Herrera (2014), uma certificação consiste na validação da conformidade de produtos, serviços, processos, sistemas ou pessoas com determinados padrões ou requisitos instituídos por uma organização.

A série de certificações ISO/IEC 25000 é organizada em: gerenciamento de qualidade, modelos de qualidade, métricas de qualidade e avaliação de qualidade. Esta última, por sua vez, possui dois importantes padrões a serem mencionados: a ISO/IEC 25012 e a ISO/IEC 25024 (GUALO et al., 2021).

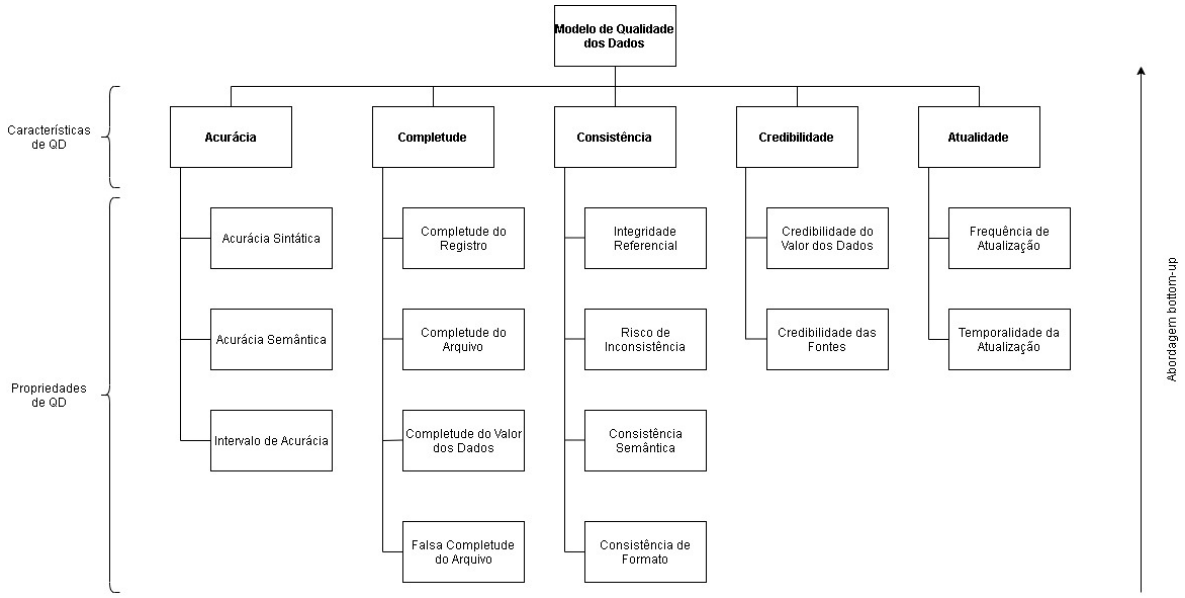
A certificação de qualidade de dados ISO/IEC 25012 é responsável pelos padrões de qualidade na indústria, providenciando uma classificação para as características que a definem. Esta, por sua vez, pode ser inerente ou dependente do sistema, sendo esta dedicada ao estudo de como a qualidade de dados é alcançada e preservada em um sistema de informação, enquanto aquela ao grau em que os dados possuem capacidade de satisfazer as necessidades implícitas dos dados (GUALO et al., 2021).

Sabendo disso, devido à diferença de natureza tecnológica das várias bases de dados, procurou-se trabalhar apenas com as características intrínsecas dos dados, como forma de uniformizar todo e qualquer processo de avaliação da qualidade dos dados, como aquelas definidas pela ISO/IEC 25012: acurácia, completude, consistência, credibilidade e atualidade. (GUALO et al., 2021).

Para que as avaliações incluíssem também as características dependentes dos sistemas de informação, criou-se a ISO/IEC 25024, a qual reúne aquelas intrínsecas definidas pela ISO/IEC 25012 mais a definição de "propriedade de qualidade"(GUALO et al., 2021). Esta, por sua vez, consiste em avaliar certos aspectos da qualidade dos dados em um repositório (GARCÍA, 2017).

Sendo assim, a Figura 2.6 ilustra como as características de qualidade se associam às propriedades de qualidade na elaboração de um modelo de qualidade dos dados segundo a ISO/IEC 25024 para avaliar a qualidade de um determinado repositório.

Figura 2.6: Características de qualidade dos dados e suas propriedades segundo a ISO/IEC 25024.



Fonte: Adaptado de Gualo et al. (2021).

Logo, por meio de uma abordagem *bottom-up*, ou seja, iniciando-se a partir das propriedades de qualidade em direção às características de qualidade, construiu-se um modelo de qualidade de dados considerando não só os dados de forma intrínseca, mas também o sistema em questão.

2.4.1 Forma de avaliação

Como forma de avaliar a qualidade dos dados contidos no repositório de uma determinada organização, García (2017), GUALO et al. (2021) e a ISO/IEC (2011) definem uma sequência de etapas que devem ser seguidas para que se obtenha a certificação de qualidade dos dados. A seguir encontram-se os cinco passos a serem seguidos no processo de avaliação:

1. Definição dos requisitos para a avaliação de qualidade dos dados;

2. Especificação de como ocorrerá a avaliação;

3. Planejamento das atividades de avaliação;

4. Execução da avaliação da qualidade dos dados;

5. Conclusão da avaliação.

Primeiramente, estabelece-se o propósito da avaliação da qualidade dos dados por meio de reuniões entre a equipe de avaliação do laboratório credenciado e os membros da organização, realizando também a introdução do modelo de qualidade de dados e as métricas utilizadas pela equipe de avaliação aos membros da organização, com o intuito de selecionar as características de qualidade de dados que mais se adéquam aos requisitos. Em seguida, obtém-se os requerimentos, identificam-se os dados alvos a serem alcançados e o rigor a ser definido no processo de avaliação. No final desta etapa, é obtida uma cópia estática do repositório, um documento especificando as regras de negócio juntamente com os requisitos de qualidade dos dados que o repositório deve atender, e quais características de qualidade dos dados serão avaliadas (GUALO et al., 2021).

Durante a etapa de especificação, são selecionadas as métricas de qualidade dos dados e os critérios de decisão, tanto para as medidas utilizadas, quanto para a avaliação, obtendo-se ao final os detalhes que levarão o processo de avaliação adiante (GUALO et al., 2021).

A terceira etapa consiste em afunilar o escopo da avaliação, estabelecendo detalhes do plano de atividades a ser executado. Feito isso, a penúltima etapa destina-se à execução do que foi proposto, de modo a satisfazer às métricas descritas no passo 2 e determinar o valor e o nível de qualidade para as propriedades e as características de qualidade, procurando identificar as possíveis forças e fraquezas nos dados do repositório (GUALO et al., 2021). Para isso, Rodríguez, Oviedo e Piattini (2016) utilizam conceitos como o de níveis, intervalos de qualidade e funções de perfilamento. Por fim, esta etapa passa pelos seguintes estágios até a sua conclusão:

5.1 Criação dos *scripts* de avaliação, os quais são utilizados para verificação das regras de negócio definidas inicialmente;

5.2 Execução dos *scripts* de avaliação para se obter as forças e fraquezas de cada propriedade de qualidade;

5.3 Produção do valor de qualidade para as propriedades de qualidade dos dados. As medidas obtidas anteriormente são utilizadas para calcular o valor de qualidade das propriedades, aplicando-lhes uma função de avaliação;

5.4 Derivação do nível de qualidade para as propriedades de qualidade a partir do valor de qualidade, atribuindo um nível de qualidade de acordo com o intervalo do valor de qualidade, por exemplo;

5.5 Determinação do nível de qualidade para as características de qualidade de dados selecionadas, calculada por meio da agregação dos níveis de qualidade das propriedades de qualidade dos dados correspondentes. A agregação é realizada em cada característica por meio da utilização de uma função de perfilamento (RODRÍGUEZ; OVIEDO; PIATTINI, 2016).

Finalmente, a última etapa no processo de avaliação é a conclusão de fato, a qual ocorre elaborando-se um relatório detalhado contendo as informações referentes aos níveis de qualidade alcançados para as características de qualidade dos dados selecionados, além dos valores obtidos para as propriedades de qualidade dos dados correspondentes. Ao final, além do relatório de avaliação, pode ainda existir um relatório com sugestões de melhorias, caso seja necessário, bem como um relatório final que descreva a revogação dos privilégios de acesso aos dados do repositório por parte da equipe avaliadora (GUALO et al., 2021).

2.5 Considerações Finais

Há uma vasta lista de algoritmos de classificação disponíveis na literatura que podem ser utilizados para determinado fim, conforme foi mostrado por Vasa (2016), Thangaraj e Sivakami (2018), Mohammed, Khan e Bashier (2016). Independente de qual for o algoritmo

escolhido, recomenda-se cautela no processo de classificação. Importante que seja visto de forma abrangente, uma vez que a qualidade dos dados pode impactar nas propriedades e características de qualidade de uma classificação (BLAKE; MANGIAMELI, 2011).

Neste sentido, utilizando as principais dimensões de qualidade para avaliação dos dados na indústria segundo a série de certificações ISO/IEC 25000 (ISO/IEC, 2011), juntamente com a seleção de um algoritmo que atenda às necessidades, seria possível obter melhores resultados na classificação de conteúdo textual.

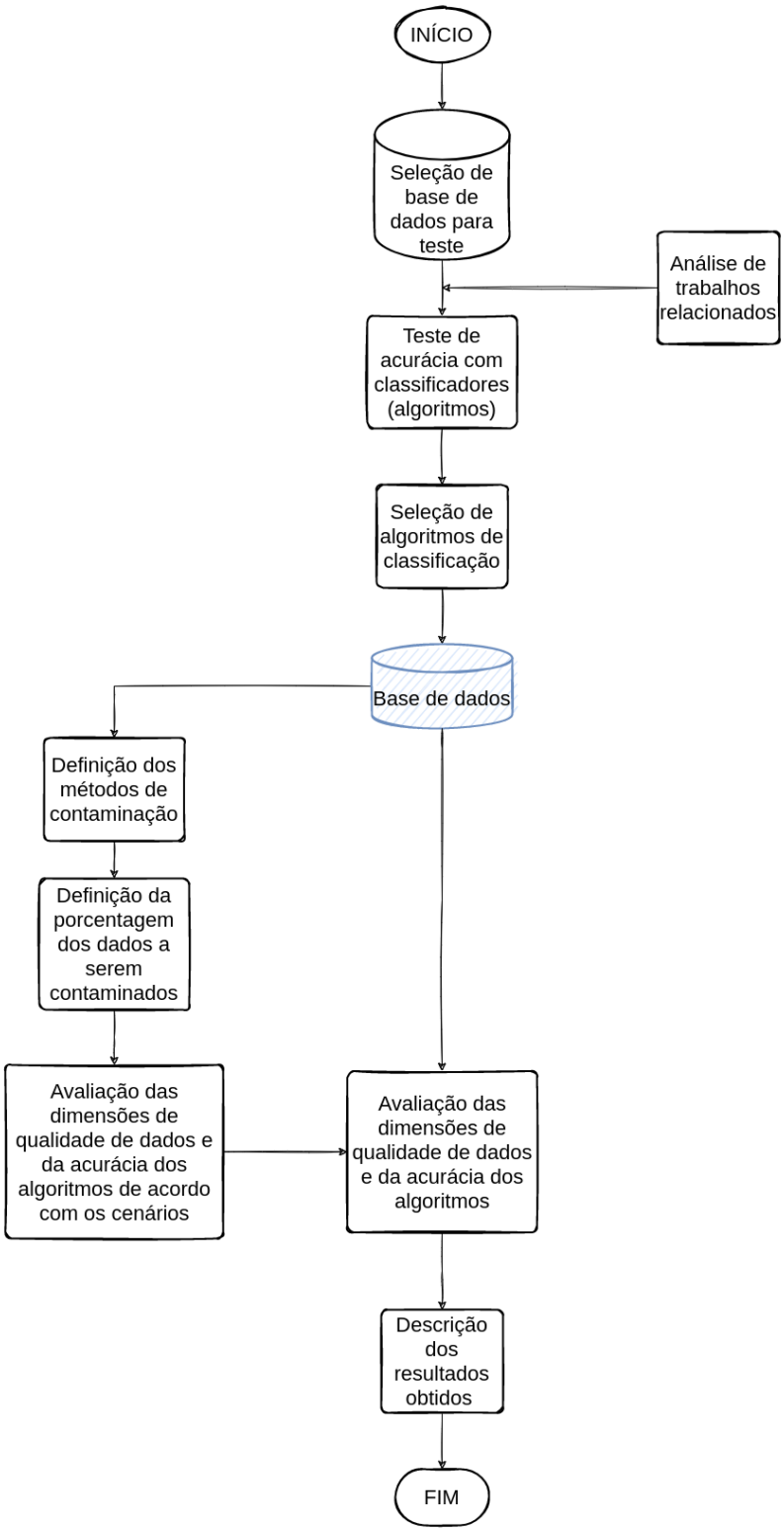
No capítulo seguinte será apresentada a metodologia utilizada no presente trabalho, buscando estabelecer critérios de comparação dos algoritmos de classificação de conteúdo textual após avaliação prévia da influência de ruídos que afetam a qualidade dos dados e, consequentemente, as principais dimensões de qualidade.

Capítulo 3

Metodologia

No capítulo contém a descrição do procedimento metodológico do trabalho. Para melhor compreensão da sequência de etapas, construiu-se um fluxograma, ilustrado pela Figura 3.1, o qual permite deixar mais evidente o fluxo de desenvolvimento do trabalho. Na Seção 3.1, foi descrito o conjunto de passos adotados para a seleção dos algoritmos de classificação considerados posteriormente. Em seguida, na Seção 3.2, foi detalhado o processo de construção e contaminação da base de dados que irá simular deficiências dos dados armazenados em um repositório de conteúdo textual. A sequência de avaliação das dimensões de qualidade de dados e da acurácia dos algoritmos, bem como da descrição das etapas até a classificação estão descritas na Seção 3.3, juntamente com as considerações finais.

Figura 3.1: Fluxograma detalhado da metodologia de implementação.



Fonte: Elaborado pelo autor.

3.1 Seleção dos algoritmos de classificação

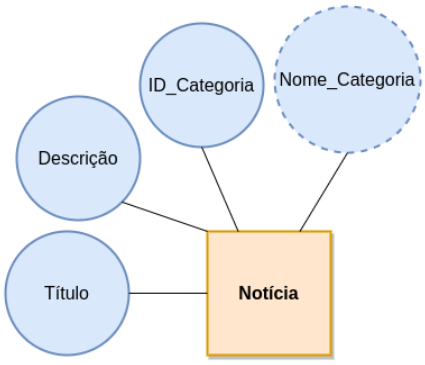
Um estudo preliminar foi realizado com intuito de selecionar os melhores algoritmos de classificação de informações textuais. Para tanto, utilizou-se uma base de teste contendo inicialmente 120.000 registros de notícias, divididas em quatro categorias distintas: "Ciência e Tecnologia", "Mundo", "Esportes"e "Negócios", e quatro atributos: título da notícia, descrição da notícia, identificador da categoria e nome da categoria, sendo este último gerado a partir dos identificadores, conforme mostra a Figura 3.2. A partir desta base, procurou-se utilizar o atributo de descrição da notícia para a classificação. Além disso, por motivos de limitação de *hardware*, realizou-se uma amostragem de 10% dos dados para o processo de classificação. A Tabela 3.1 apresenta as especificações gerais da plataforma utilizada para os testes.

Tabela 3.1: Especificações da máquina utilizada para a realização dos testes.

Componente	Descrição
Processador	Intel(R) Core(TM) i5-5200U 2,20 GHz com 4 núcleos
Memória RAM	8GB
Sistema Operacional	Ubuntu 20.04.2 LTS de arquitetura 64 bits
GPU	GeForce 940M com 1072 MHz de processamento e 2GB

Fonte: Elaborado pelo autor.

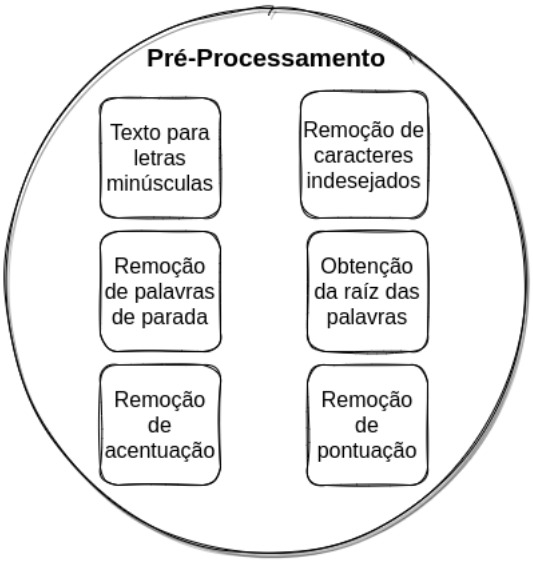
Figura 3.2: Entidade **Notícia** e seus atributos correspondentes.



Fonte: Elaborado pelo autor.

Dessa maneira, seguindo o procedimento denotado pela Figura 2.1, foi realizada uma etapa de pré-processamento a fim de remover elementos que pudessem interferir na classificação. Assim, a Figura 3.3 mostra os métodos utilizados para a remoção de ruídos:

Figura 3.3: Medidas adotadas para o pré-processamento dos textos.



Fonte: Elaborado pelo autor.

Assim, as técnicas de passagem do texto para letras minúsculas, bem como as de remoção de caracteres indesejados, acentuações e pontuações, visaram remover características irrelevantes ou redundantes as quais poderiam comprometer o desempenho e a acurácia dos algoritmos de classificação (BAHARUDIN et al., 2010). Além disso, foi realizada a remoção de palavras de parada, tais como preposições e outras cuja relevância para a classificação é nula, bem como a obtenção da raiz das palavras por meio do processo de lematização, o qual substitui ou remove o sufixo de uma palavra para obter sua forma básica (KOWSARI et al., 2019).

Em seguida, utilizou-se o método denominado Frequência dos Termos - Frequência Inversa dos Documentos, ou *Term Frequency - Inverse Document Frequency* (TF-IDF) em inglês, para a extração de características, cujo objetivo é proporcionar redução de dimensionalidade no espaço de características. Esta técnica é conhecida no campo da mineração de textos,

sendo utilizada como um fator de ponderação na importância das palavras nos documentos e podendo até ser comparada com outros métodos mais recentes em termos de performance (KADHIM, 2019).

Kowsari et al. (2019) relata que Jones (1972) foi quem propôs o método de Frequência Inversa dos Documentos, o qual foi utilizado em conjunto com a Frequência dos Termos para reduzir o efeito de palavras comuns que aparecem implicitamente no *corpus* (conjunto de palavras). O princípio de funcionamento da Frequência Inversa dos Documentos consiste em atribuir pesos maiores para palavras com baixa ou alta frequência presentes nos documentos. Assim, a determinação do peso de um termo em um texto pode ser descrito pela seguinte equação matemática:

$$P(d,t) = FT(d,t) * log(\frac{N}{fd(t)}). \tag{3.1}$$

Nessa equação, N representa o número de documentos e fd(t) o número de registros contendo o termo t no *corpus*. Apesar de relativamente boa em termos de desempenho, a técnica sofre algumas limitações, como a incapacidade de lidar com termos similares devido às palavras serem armazenadas como um índice de forma independente. Para lidar com isso, outros métodos têm sido desenvolvidos (KOWSARI et al., 2019).

Dentre os mencionados no Capítulo 2, foram selecionados seis classificadores para a validação do modelo criado: Floresta Aleatória, Máquina de Vetores de Suporte, *Naive Bayes* Multinomial, Regressão Logística, k-Vizinhos Mais Próximos e o *Perceptron* Multicamadas.

A escolha do conjunto de classificadores utilizados para teste foi influenciada por trabalhos relacionados (KOWSARI et al., 2019; THANGARAJ; SIVAKAMI, 2018; VASA, 2016; BRINDHA; PRABHA; SUKUMARAN, 2016) os quais analisaram o uso de algoritmos supervisionados para a classificação de textos, como aqueles expressos na Tabela 2.1. Dessa maneira, foi possível compreender os algoritmos no que se refere ao funcionamento, as vantagens e as desvantagens na utilização de cada um estabelecendo um comparativo. Vale ressaltar que trabalhos mais recentes, como a revisão da literatura apresentada por Minaee et

al. (2021), apontam para a utilização de modelos de aprendizado de máquina profundos, os quais são capazes de superar os modelos tradicionais de aprendizado de máquina em tarefas de processamento de linguagem natural, tais como: análise de sentimentos, categorização de notícias, entre outras. Dito isso, os modelos de aprendizado de máquina profundos não são explorados no contexto deste trabalho.

Após a seleção dos classificadores, utilizou-se a técnica de validação cruzada *k-folds*, a qual reúne uma parte dos dados disponíveis para se ajustar ao modelo e outra para a realização de testes. Assim, os dados são divididos em k partes iguais, e em cada uma das k iterações, uma das partes é validada e as outras são utilizadas para treinamento, combinando ao final as k estimativas de erro na predição (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Dessa maneira, realizou-se a validação cruzada três vezes para cada um dos seis algoritmos, obtendo-se os resultados de acurácia de cada iteração. A partir da análise dos resultados obtidos da validação cruzada juntamente com a média de acurácia resultante, foram selecionados apenas os algoritmos que obtiveram resultados satisfatórios.

3.1.1 Desempenho dos algoritmos de classificação no estudo preliminar

Após a obtenção e amostragem de 10% dos dados provenientes da base de testes, foi realizado o pré-processamento e, em seguida, a extração de características por meio da aplicação do TF-IDF, ignorando termos com frequência de documento menor que três e extraindo tanto uni-gramas como bigramas a partir dos textos. Dessa forma, a partir das 12000 amostras selecionadas, foi obtido um vetor de 17962 posições para compor o vetor de características. Por fim, aplicou-se a validação cruzada três vezes para os algoritmos.

Em termos de desempenho, foi necessário um tempo de 45 minutos e 6 segundos para a classificação. Foi realizada, então, a média dos três *folds* para cada um dos algoritmos, e o resultado apresentado por meio da Tabela 3.2.

Diante do comparativo exposto acima, Naive Bayes Multinomial, Máquina de Vetores de Suporte e Regressão Logística foram os algoritmos selecionados, considerando as maiores

Tabela 3.2: Desempenho nos testes dos três *folds*.

	Acurácia (%)	
Classificador	Média	Desvio Padrão
Naive Bayes Multinomial	**88,94**	**3,82E-04**
Regressão Logística	**88,41**	**4,06E-03**
Máquina de Vetores de Suporte	**88,36**	**4,86E-03**
Perceptron Multicamadas	86,98	3,97E-03
k-Vizinhos Mais Próximos	85,27	4,59E-03
Floresta Aleatória	84,07	4,51E-03

Fonte: Elaborado pelo autor.

médias e menores desvios.

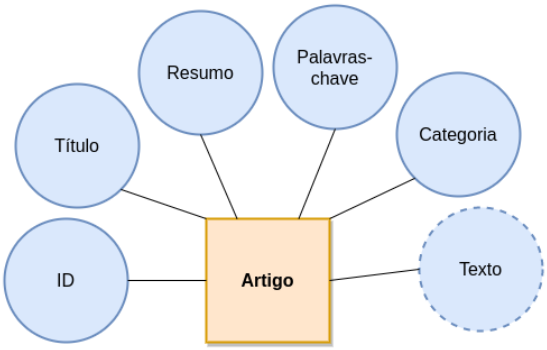
Em suma, a etapa de seleção criteriosa dos algoritmos de classificação foi um passo importante do procedimento metodológico para realizar, uma vez que foi fundamental neste estudo preliminar para economizar tempo e recursos computacionais na base de dados de avaliação construída posteriormente.

3.2 Base de dados de avaliação

A base de dados de avaliação foi construída por meio da coleta manual de artigos científicos a partir da base de dados da SciELO (2021), um repositório contendo trabalhos desenvolvidos nas mais diversas áreas de atuação e de vários países. A base de dados é composta por seis atributos: identificador, título, resumo, palavras-chave e categoria do artigo, além de um novo atributo, denominado "Texto", criado a partir dos já existentes e que reúne as informações presentes no título, resumo e nas palavras-chave, conforme ilustra a Figura 3.4.

Para a classificação, nove possibilidades de classes foram fornecidas, conforme as áreas temáticas da própria SciELO: "Ciências da Saúde", "Ciências Humanas", "Ciências Sociais Aplicadas", "Ciências Agrárias", "Ciências Biológicas", "Ciências Exatas e da Terra", "Engenharias", "Linguística, Letras e Artes"e "Multidisciplinar". Dessa forma, tais áreas configuraram os valores possíveis para o atributo "Categoria". Por fim, a classificação foi realizada a

Figura 3.4: Entidade **Artigo** e seus atributos correspondentes.

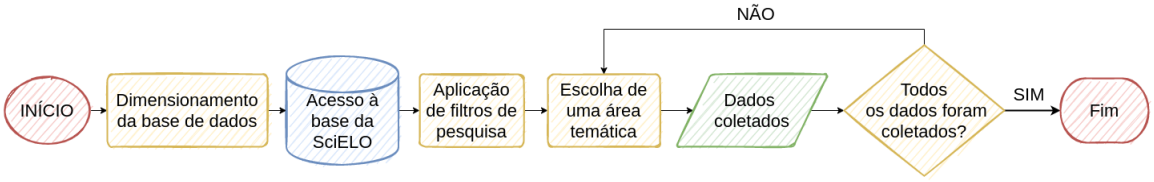


Fonte: Elaborado pelo autor.

partir dos dados textuais presentes no atributo "Texto", para que mais informações pudessem ser extraídas de maneira a melhorar os resultados das classificações.

O processo de coleta dos dados para a criação do repositório seguiu os passos descritos na Figura 3.5. Dessa maneira, foi possível garantir que o procedimento fosse padronizado de forma a evitar possíveis inconsistências.

Figura 3.5: Procedimento de construção da base de dados científicos.



Fonte: Elaborado pelo autor.

Ao todo, 1800 artigos científicos foram coletados para compor a base de dados científicos (BDC), sendo estes divididos igualmente entre as nove categorias possíveis, totalizando 200 por área temática. Em seguida, para as pesquisas na base de dados da SciELO, foram aplicados filtros específicos, tais como: apenas coleções provenientes do Brasil, idioma português, artigo como tipo de literatura e áreas temáticas variadas da *Web Of Science*, as quais foram utilizadas para denotar as sub-áreas existentes em cada uma das nove grandes áreas. Dessa forma, foi possível prover variabilidade nos dados coletados, evitando introduzir possíveis

viéses no conteúdo coletado.

3.2.1 Experimento de simulação: método de contaminação dos dados

Como parte do planejamento do experimento de simulação proposto com fins de avaliar de forma comparativa os algoritmos classificadores, foi necessário primeiramente definir métodos de contaminação que possibilitassem observar os efeitos sobre as dimensões de qualidade dos dados, bem como sobre a acurácia dos algoritmos de classificação comparados. A Tabela 3.3 apresenta os métodos propostos para cada dimensão de qualidade de dados a ser avaliada por meio deles.

Tabela 3.3: Dimensão da qualidade e método de contaminação proposto para afetá-la.

Dimensão	Método
Acurácia	Adição de caracteres
Acurácia	Remoção de caracteres
Acurácia	Substituição de caracteres
Completude	Omissão de atributo
Consistência	Ambiguidade por troca de atributos

Fonte: Elaborado pelo autor.

Mais especificamente em relação aos métodos expostos na Tabela 3.3 e às dimensões de qualidade que serão avaliadas, segue um detalhamento abaixo.

- Acurácia (DQ-ACC): métodos de adição, remoção e substituição de caracteres foram definidos. A utilização destes métodos de inserção de ruídos na base de dados visou simular situações passíveis de acontecer no mundo real, como erros ortográficos, de digitação, e outras falhas que podem comprometer a qualidade dos dados armazenados. Segundo Jurafsky e Martin (2021), a frequência de erros ortográficos considerando textos escritos pode ficar entre 10-15% para consultas na rede. Assim, a função empregada para a contaminação da DQ-ACC foi implementada por meio da adoção de uma probabilidade de 10% para cada caractere contido no atributo correspondente ser afetado

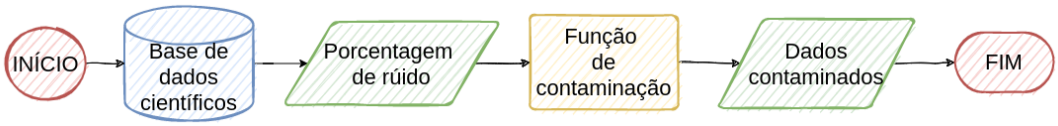
por algum método. Assim, a função inclui a opção de selecionar o método utilizado, seja este adição, remoção ou substituição de caracteres, conforme mostra a Tabela 3.3. Deste modo, propriedades como a acurácia sintática, apresentada na Figura 2.6, foram comprometidas por meio da aplicação dessas técnicas.

- Completude (DQ-COMP): adotou-se o método de omissão de atributo, o qual implica em apagar os dados contidos em um atributo aleatório do registro contaminado. Dessa maneira, a propriedade de completude do registro, contida na Figura 2.6, seria afetada. Assim, situações em que um usuário deixa de preencher um campo obrigatório, por exemplo, puderam ser contempladas no experimento de simulação por meio deste método.

- Consistência (DQ-CONS): utilizou-se a abordagem de trocar o conteúdo textual de um atributo de uma determinada instância pelo texto de um atributo aleatório de uma instância de outra classe, produzindo inconsistências e ambiguidades na base de dados. Assim, a propriedade de integridade referencial contida na Figura 2.6 seria violada pela contaminação.

Dessa maneira, após a base de dados ter sido construída, procedeu-se para a inserção dos ruídos por meio do uso de uma função desenvolvida para a contaminação, conforme mostra a Figura 3.6.

Figura 3.6: Sequência de etapas para a inserção de ruídos na base de dados científicos.



Fonte: Elaborado pelo autor.

3.2.2 Descrição das condições experimentais

As condições experimentais para composição das bases de dados contaminadas foram estabelecidas nesta etapa. O experimento foi planejado para avaliar cinco condições experimentais, a saber:

1. Grupo controle (sem adição de ruído);
2. Grupo ruído I (com adição de 10% de ruído);
3. Grupo ruído II (com adição de 20% de ruído);
4. Grupo ruído III (com adição de 30% de ruído);
5. Grupo ruído IV (com adição de 40% de ruído).

Os percentuais de ruído foram aplicados de forma aleatória (randômica) em todos os métodos de contaminação conforme o percentual estabelecido na condição experimental.

Um exemplo prático seria considerar a condição experimental de 20% dos dados contaminados. Isto significaria que 360 dos 1800 artigos presentes na BDC seriam afetados por algum método definido a partir das dimensões de qualidade de dados selecionadas, como mostra a Tabela 3.3. Assim, levando em consideração a DQ-ACC, o método de adição de caracteres e o atributo ”Resumo”, o resultado obtido seria algo próximo do que é apresentado pela Tabela 3.4, considerando a probabilidade de 10% de adição de um novo caractere aleatório para cada um analisado ao longo do texto.

Tabela 3.4: Exemplo de contaminação da DQ-ACC por meio do método de adição de caracteres.

Texto sem contaminação	Texto contaminado
’As variações das propriedades óticas dos aerossóis podem interferir nos processos de transferência de energia entre a atmosfera e a superfície terrestre. (...) ’	’AsF varTiTaSções das propriredades óticas Kdos aerossóis podemT intetrbferir nos processQofs de transferência de enhergia entre a adtmosfera e aI superfífciie tGerrestre. (...) ’

Fonte: Elaborado pelo autor.

Como a ideia é avaliar o impacto da qualidade dos dados na acurácia do algoritmo de classificação, para cada condição experimental, foi calculada a acurácia do algoritmo de classificação como a principal variável resposta para avaliação dos resultados.

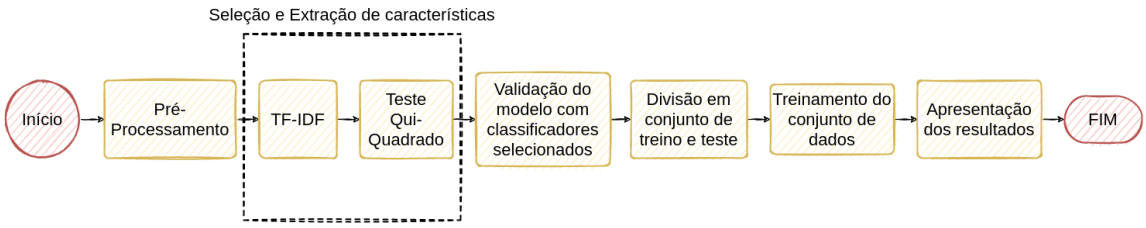
3.3 Classificação e análise dos resultados

Para a etapa de classificação e avaliação dos resultados, tanto em relação às dimensões de qualidade, quanto à acurácia dos algoritmos, dois cenários foram propostos: um com a base contendo ruídos e outro com a base "limpa", ou sem contaminação, exatamente como ilustrado na Figura 3.1. Dessa forma, procurou-se compreender os efeitos da contaminação sobre as métricas consideradas.

3.3.1 Processo de classificação

Para a sequência metodológica utilizada até a classificação, seguiu-se o fluxo descrito pela Figura 2.1. A Figura 3.7 detalha um pouco mais o fluxo de trabalho desta etapa.

Figura 3.7: Sequência de etapas seguidas durante o processo de classificação.



Fonte: Elaborado pelo autor.

Sendo assim, primeiramente foi realizada uma etapa de pré-processamento nos registros da base. Para tal, utilizaram-se as mesmas técnicas adotadas na base de dados de teste, ilustradas pela Figura 3.3. Uma diferença, porém, está no método usado para a extração da raiz das palavras: enquanto na base de dados de teste foi utilizada a técnica de lematização, na BDC usou-se a derivação, cujo funcionamento consiste em modificar as palavras a fim de obter variantes por meio de processos linguísticos, como a afixação (KOWSARI et al., 2019).

Em seguida, visando a redução da dimensionalidade do espaço de características, foram utilizadas técnicas de seleção e extração das mesmas. A extração de características foi realizada por meio do mesmo método empregado na base de teste, o TF-IDF, enquanto que para a seleção se usou o Teste Qui-Quadrado. Este, por sua vez, baseia-se em examinar a independência de dois eventos, neste caso as classes e os termos em questão (SHAH; PATEL, 2016).

$$\text{Ho: As classes não estão associadas aos termos } p(AB) = p(A)p(B) \tag{3.2}$$

Para testar a hipótese Ho pelo teste qui-quadrado, deve-se considerar a seguinte estatística de teste:

$$X^2(t,c) = \frac{\sum_{t\in(0,1)} \sum_{c\in(0,1)} (O_{t,c} - E_{t,c})^2}{E_{t,c}}. \tag{3.3}$$

A partir dela, tem-se que O representa a frequência observada e E a frequência esperada para cada estado da classe c e do termo t. Dessa maneira, por meio do teste qui-quadrado, é possível mensurar o quanto as contagens esperadas, denotadas por E, e as observadas, representadas por O, divergem entre si (SHAH; PATEL, 2016).

A validação do modelo criado com os classificadores selecionados é feita a partir da filtragem inicial realizada por meio da classificação na base de testes e da análise de trabalhos correlatos. Para isso, utiliza-se novamente a técnica de validação cruzada *k-folds*, porém desta vez com 10 *folds* para cada um dos algoritmos, e os resultados são plotados em um diagrama de caixas para efeitos de comparação.

Além disso, como forma de descartar possíveis vieses na base de dados, utilizou-se o método de validação cruzada *leave- one-out*, o qual consiste em um tipo especial de validação cruzada no qual o número de iterações realizadas equivalem ao número de instâncias da base de dados, sendo especialmente útil quando o número de registros contidos no repositório é pequeno (WONG, 2015) já que o método também exige mais recursos computacionais para

ser implementado (MAO et al., 2012).

Feito isso, a base de dados foi dividida em um conjunto de treino e outro de teste em uma proporção adequada para proporcionar bons resultados sem causar sobreajuste dos dados, isto é, fazer com que os dados se encaixem muito bem ao modelo, mas falhem no momento de novas predições (YING, 2019). Em seguida, realizou-se o treinamento do modelo efetivamente. Por fim, os resultados foram avaliados por meio de uma matriz de confusão, a qual possibilita a comparação entre verdadeiros positivos e verdadeiros negativos em uma amostra de teste com os dados positivos e negativos previstos (SAMMUT; WEBB, 2011). A Tabela 3.5 define os termos utilizados por uma matriz de confusão. A anotação prevista recupera aquilo que foi obtido pelo algoritmo, enquanto a anotação de ouro é realizada por um humano.

Tabela 3.5: Termos utilizados na construção da matriz de confusão.

		Anotação prevista	
		Positivo	Negativo
Anotação de ouro	Positivo	Verdadeiro positivo (*vp*)	Falso negativo (*fn*)
	Negativo	Falso positivo (*fp*)	Verdadeiro negativo (*vn*)

Fonte: Adaptado de Dalianis (2018).

Outras métricas também foram obtidas, como a precisão, a revocação e a medida F (*F1-measure*, em inglês). Tanto a precisão quanto a revocação são utilizadas como medidas de performance em um sistema de recuperação (DALIANIS, 2018) e podem ser calculadas por meio das seguintes equações:

$$\text{Precisão} = \frac{vp}{vp + fp} \tag{3.4}$$

e

$$\text{Revocação} = \frac{vp}{vp + fn} \tag{3.5}$$

Nas Eq. (3.4) e (3.5), os termos utilizados podem ser encontrados por meio da matriz de

confusão apresentada pela Tabela 3.5.

Já a medida F é a média harmônica da precisão e da revocação e assume um valor no intervalo (0,1), sendo que valores mais próximos de um implicam em maior desempenho (NARASIMHAN et al., 2016). Dado um classificador h, a medida F pode ser denotada pela seguinte equação:

$$F_1(h) = \frac{2 * P(h) * R(h)}{P(h) + R(h)}. \tag{3.6}$$

Neste cálculo, *P(h)* e *R(h)* representam a precisão e a revocação aplicados sobre um classificador h, respectivamente.

Por fim, a acurácia do classificador consiste em outra métrica definida como a proporção de instâncias verdadeiras recuperadas, tanto positivas quanto negativas, entre todas as outras. Sendo assim, a acurácia pode ser expressa por uma média aritmética ponderada entre a precisão e o inverso desta (DALIANIS, 2018), conforme a equação a seguir:

$$\text{Acurácia} = \frac{vp + vn}{vp + vn + fp + fn} \tag{3.7}$$

Sendo assim, de acordo com a matriz de confusão apresentada na Tabela 3.5, a acurácia representa a razão entre as anotações verdadeiras e todas as possibilidades existentes.

Como outros trabalhos apontaram a acurácia como a métrica mais utilizada, tanto em problemas binários, quanto aqueles envolvendo múltiplas classes (CHAWLA; JAPKOWICZ; KOłCZ, 2004; GU; ZHU; CAI, 2009; HOSSIN et al., 2011; RANAWANA; PALADE, 2006), optou-se por usá-la para avaliar o desempenho da classificação. Porém, como forma de avaliar os resultados com mais confiabilidade, construiu-se também a área sob a Curva Característica de Operação do Receptor (conhecida como curva ROC), visto que este é um método que comprovadamente gera resultados melhores que a acurácia (HUANG; LING, 2005), apesar de exigir um custo computacional elevado, especialmente em problemas envolvendo múltiplas classes (HOSSIN; SULAIMAN, 2015). Para isso, foi adaptado o método cuja utilização é apropriada em problemas envolvendo duas classes para o caso multi-classes. Dessa forma,

as curvas ROC foram geradas para as classes fixando uma por vez e classificando-a contra cada uma das outras. Como resultado, tem-se no eixo horizontal do gráfico construído a Taxa de Falsos Positivos (TFP), descrita pela Eq. (3.9), a qual é equivalente ao complementar da especificidade, expressa pela Eq. (3.8), indicando a proporção em que a classe negativa foi incorretamente classificada. Em ambas as equações, *vn* representa verdadeiros negativos e *fp* falsos positivos.

$$\text{Especificidade} = \frac{vn}{vn + fp} \tag{3.8}$$

$$\text{TFP} = \frac{fp}{vn + fp} \tag{3.9}$$

Já no eixo vertical, encontra-se a sensibilidade, correspondendo à proporção da classe positiva corretamente classificada, sendo esta equivalente à revocação, denotada pela Eq. (3.5). Para avaliar os resultados, deve-se observar a área sob a curva, também conhecida como AUC, sendo que quanto maior seu valor, maior será a discriminação entre as classes positivas e negativas. Hoo, Candlish e Teare (2017) apontam que a AUC pode ser um indicador da ”acurácia” geral, pois há apenas uma área sob a curva para cada uma das curvas ROC. Apesar disso, esta ”acurácia” sofre oscilações no decorrer da curva ROC, devido às variações de especificidade e sensibilidade.

Neste capítulo foram descritos os aspectos metodológicos envolvidos na seleção e adequação de algoritmos de classificação com vistas à aplicação no contexto de uma BDC. Também foram detalhados os elementos envolvidos no experimento de simulação da contaminação da BDC construída. Para nortear todo o estudo de avaliação do impacto da qualidade dos dados na acurácia da classificação, foram introduzidos conceitos estatísticos importantes que definem critérios de avaliação baseados em medidas diagnósticas. No próximo capítulo serão apresentados os resultados da implementação da execução da metodologia aqui descrita.

Capítulo 4

Resultados

Este capítulo apresenta os resultados e análises provenientes dos testes realizados seguindo a metodologia descrita no Capítulo 3.

Os testes foram realizados por um computador seguindo as especificações descritas pela Tabela 3.1, de modo que os algoritmos de classificação selecionados - Máquina de Vetores de Suporte (MVS), Naive Bayes Multinomial (NBM) e Regressão Logística (RL) - foram avaliados para cada uma das cinco condições experimentais pré definidas (Grupo controle; Grupo ruído I; Grupo ruído II; Grupo ruído III; Grupo ruído IV).

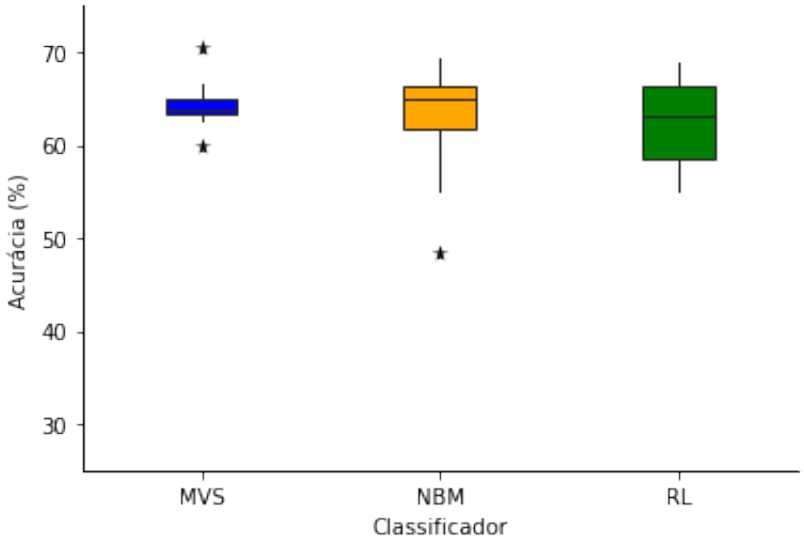
As dimensões de qualidade de dados Acurácia (DQ-ACC), Completude (DQ-COMP) e Consistência (DQ-CONS) foram quantificadas em cada teste, levando em consideração cada um dos atributos das instâncias presentes na BDC, a saber: o título, resumo, palavras-chave e o texto completo, (vide Figura 3.4). Importante salientar que a quantificação das dimensões foi efetuada com base nos métodos descritos na Tabela 3.3.

4.1 Comparação entre os algoritmos

Após as etapas de pré-processamento, extração e seleção de características, como descritas na Figura 3.7, obteve-se o modelo de classificação, conforme mostra a Figura 2.1, o qual pôde ser avaliado por meio do uso de métodos de validação. Para isso, utilizou-se inicial-

mente a validação cruzada com 10 *folds*, como descrito na Seção 3.3.1, sendo esse um valor comumente utilizado para este tipo de validação (GARCíA; LUENGO; HERRERA, 2014). Os resultados dos *folds* são ilustrados na forma de um diagrama de caixas pela Figura 4.1. Além disso, como forma de quantificar, a média aritmética e o desvio padrão dos 10 *folds* são descritos pela Tabela 4.1.

Figura 4.1: Diagrama de caixas da acurácia obtida pelos classificadores por meio da validação cruzada de 10 *folds*.



Fonte: Elaborado pelo autor.

Tabela 4.1: Acurácia média e desvio padrão dos 10 folds da validação cruzada para cada classificador.

	Acurácia (%)	
Classificador	Média	Desvio Padrão
MVS	64,39	2,75
RL	62,61	4,89
NBM	62,56	6,40

Fonte: Elaborado pelo autor.

O algoritmo MVS obteve 64,39% de acurácia média com desvio padrão de 2,75, enquanto para os classificadores NBM e RL alcançaram 62,56% e 62,61% de acurácia média, com des-

vios padrão de 4,89 e 6,40 respectivamente. Tal resultado sugere que o MVS seja mais estável uma vez que o desvio padrão foi menor dos três. Na Figura 4.1 verifica-se a presença de dois pontos discrepantes para o MVS que não afetaram sua estabilidade na classificação. Destaca-se ainda que o classificador NBM apresentou o maior desvio padrão, aproximadamente 2,3 vezes maior que o MVS, indicando sua instabilidade na classificação, possivelmente devido à presença de um ponto discrepante representado por um *fold* com acurácia abaixo de 50%.

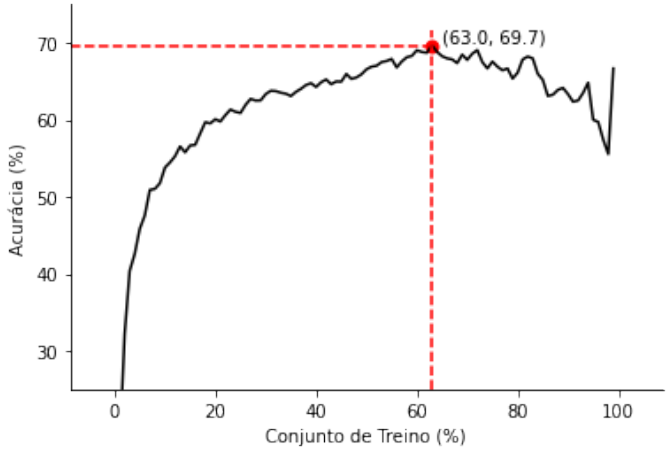
Prosseguindo com a comparação entre os classificadores, foi realizada a divisão do conjunto de dados em duas partes: uma para treinamento e outra para testes. Dito isso, não houve a inclusão de um conjunto à parte destinado à validação. A divisão, ou *split*, foi definida de acordo com as melhores divisões possíveis encontradas para cada um dos algoritmos, possibilitando encontrar a melhor acurácia dentro dos cenários possíveis, conforme ilustra a Figura 4.2.

Como se pode observar, o algoritmo MVS obteve um resultado superior aos demais classificadores quando o conjunto de treino correspondeu a 63% dos dados da BDC (37% dos dados utilizados para testes). Enquanto MVS atingiu a acurácia máxima de 69,7% (Figura 4.2a), os outros dois algoritmos alcançaram melhores resultados quando o conjunto de treino equivaleu a 82% dos dados da base (18% dos dados utilizados para testes). Não superaram o MVS na acurácia máxima.

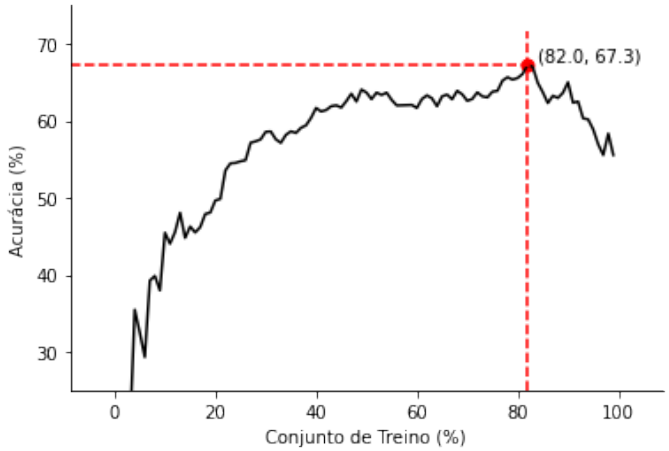
Utilizando-se dos resultados encontrados por meio da validação cruzada com 10 *folds* a partir da Tabela 4.1 e das divisões ótimas obtidas a partir da Figura 4.2, aplicou-se também a validação cruzada *leave-one-out* nos classificadores para compará-los. Assim, a Tabela 4.2 reúne os resultados obtidos em termos de acurácia para os três métodos de validação descritos, bem como o tempo de execução despendido pelo processador para efetuá-los.

A partir dos resultados obtidos, observa-se que o algoritmo MVS alcançou desfechos mais positivos em relação à acurácia obtida por meio dos métodos de validação quando comparado aos outros classificadores. Além disso, nota-se que, apesar dos resultados inferiores, o NBM despendeu menos recursos computacionais que os outros dois algoritmos em todos os méto-

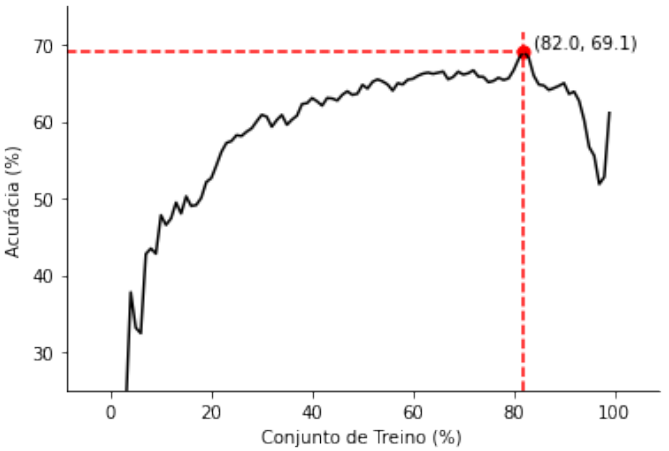
Figura 4.2: Divisões do conjunto de treino e respectivas acurácias para os classificadores.



(a) Máquina de Vetores de Suporte.



(b) Naive Bayes Multinomial.



(c) Regressão Logística.

Fonte: Elaborado pelo autor.

Tabela 4.2: Acurácia e tempo de execução do processador utilizando diferentes métodos de validação para os classificadores.

Classificador	Método de validação	Acurácia (%)	Tempo de execução (s)
MVS	Validação cruzada *leave-one-out*	69,6	573
	Divisão do conjunto de dados	69,7	159 x 10^{-3}
	Validação cruzada 10 *folds*	64,4	3
NBM	Validação cruzada *leave-one-out*	64,8	274
	Divisão do conjunto de dados	67,3	139 x 10^{-3}
	Validação cruzada 10 *folds*	62,6	2
RL	Validação cruzada *leave-one-out*	66,3	570
	Divisão do conjunto de dados	69,1	152 x 10^{-3}
	Validação cruzada 10 *folds*	62,6	3

Fonte: Elaborado pelo autor.

dos de validação utilizados.

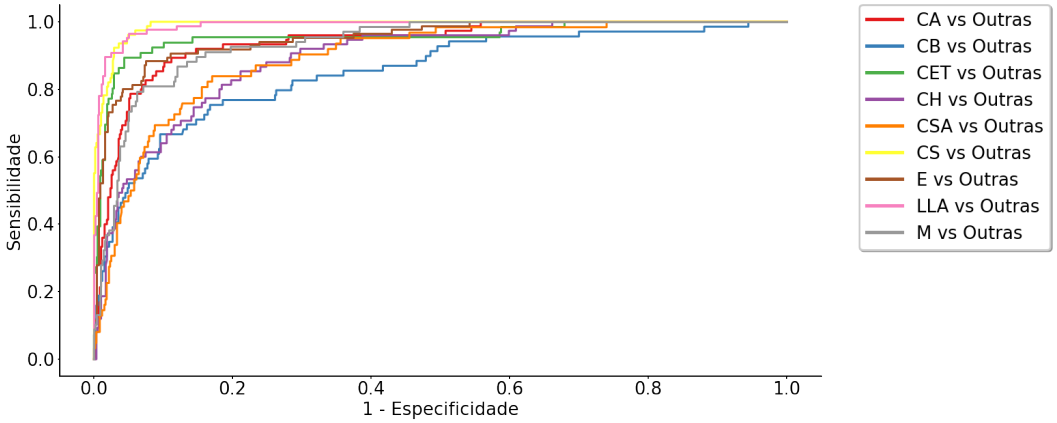
A fim de explorar uma possível influência da área de conhecimento do trabalho científico contido na BDC, procedeu-se com a comparação dos algoritmos a partir da construção da curva ROC comparativa, isto é, uma curva para cada uma das áreas de conhecimento. Dessa forma, pôde-se obter resultados mais confiáveis, pois utilizar a acurácia revelou-se muito limitada para obtenção de conclusões mais assertivas neste ponto. Cabe salientar que a curva ROC ajuda a produzir resultados mais elucidativos (YANG, 1999).

4.1.1 Análise da influência das áreas de conhecimento baseada na curva ROC

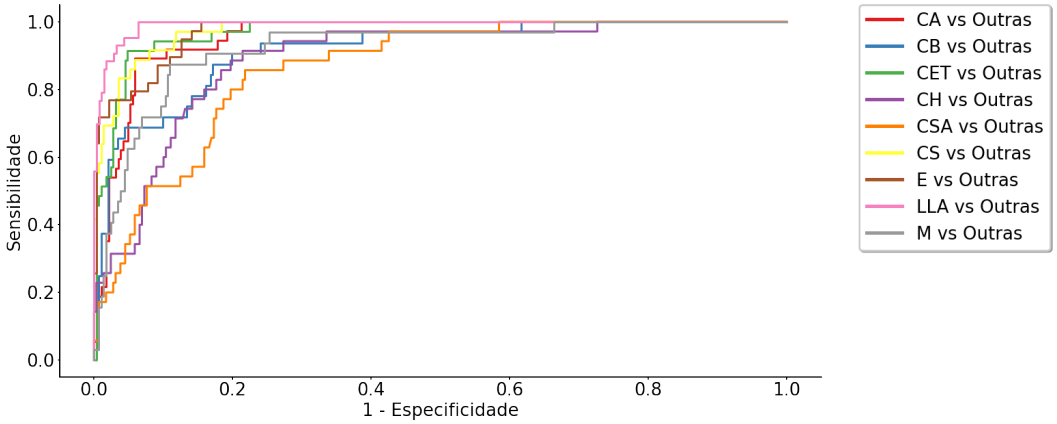
Como forma de avaliar também a diferença na classificação dos artigos da BDC em relação às áreas de conhecimento, foram construídas as curvas ROC para cada um dos classificadores, de acordo com a melhor divisão possível obtida a partir da Figura 4.2. Os resultados são expressos pela Figura 4.3. Tendo em vista o contexto multi-classes em que o problema está inserido, as curvas foram obtidas por meio da fixação de cada uma das classes, ou áreas do conhecimento, realizando a classificação contra as outras oito, rotuladas como ”Outras”.

De modo geral, artigos científicos pertencentes às áreas de: ”Ciências Sociais Aplicadas”

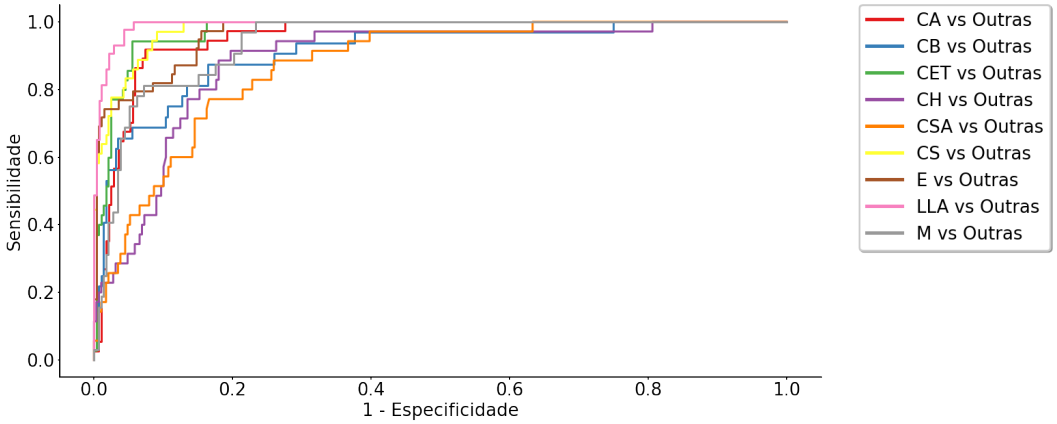
Figura 4.3: Curva ROC obtida para as diferentes áreas do conhecimento de acordo com cada classificador.



(a) Máquina de Vetores de Suporte.



(b) Naive Bayes Multinomial.



(c) Regressão Logística.

Fonte: Elaborado pelo autor.

(CSA), ”Ciências Humanas” (CH) e ”Ciências Biológicas” (CB) não foram bem classificadas de forma geral pelos três algoritmos. Já para a área de ”Linguística, Letras e Artes” (LLA) percebeu-se destaque positivo.

Mais especificamente, o algoritmo de MVS se destacou na classificação de artigos da área de ”Ciências da Saúde” (CS). Entretanto, obteve resultados não satisfatórios em relação à área de ”Ciências Biológicas”. Em contrapartida, NBM e RL alcançaram desempenho não satisfatório na classificação dos artigos pertencentes à área de ”Ciências Sociais Aplicadas”.

4.1.2 Análise comparativa

Para a avaliação da inserção dos ruídos na base foram construídas três funções de contaminação, uma para cada dimensão de qualidade já mencionada. Feito isso, os quatro atributos contendo dados textuais - título, resumo, palavras-chave e texto - foram submetido às classificações utilizando um dos três algoritmos e sob diferentes condições experimentais, como definido na Seção 3.2.2. Para garantir resultados mais consistentes, as classificações foram realizadas utilizando validação cruzada com 10 *folds* e as execuções repetidas 10 vezes, devido ao caráter aleatório embutido nas funções de contaminação.

Como foi mencionado na Seção 1.3, existe uma relação direta entre a baixa qualidade de dados e problemas organizacionais (GUALO et al., 2021). No contexto atual, a disponibilidade de dados, bem como sua heterogeneidade, são cada vez mais presentes devido à era do *Big Data* (CAI; ZHU, 2015). Assim, cabe avaliar o impacto dos ruídos que simulam situações cujas grandes bases de dados podem estar suscetíveis, sobretudo se existe a presença de dados propensos a serem atribuídos incorretamente, como os textuais (SUBRAMANIAM et al., 2009; VINCIARELLI, 2005).

Para a contaminação, cada uma das dimensões de qualidade de dados foram afetadas segundo os métodos definidos pela Tabela 3.3, e de acordo com o que foi detalhado na Seção 3.2.1. Os resultados obtidos são expressos por meio dos diagramas de caixas apresentados nas Figuras 4.4, 4.5 e 4.6, as quais se referem a DQ-ACC (adição de caracteres), DQ-COMP

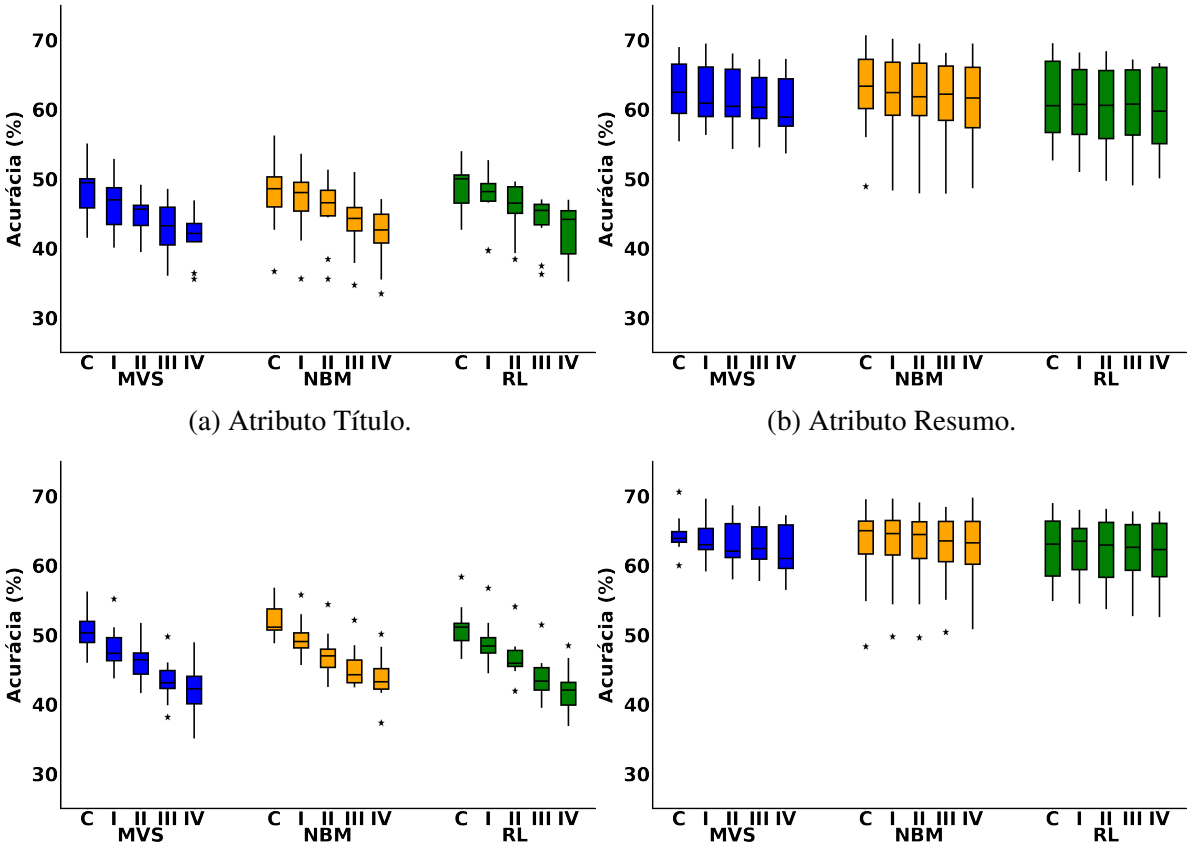
e DQ-CONS, respectivamente.

O método de análise da variância de dois fatores, mais conhecido como *Two-way ANOVA* foi aplicado para comparar os grupos que estabelecem as condições experimentais quanto às médias de acurácia. Analisou-se ainda as diferenças entre os classificadores e a interação entre grupo e algoritmo. As análises foram efetuadas para cada atributo, tomando a acurácia como a resposta de interesse. Os resultados mais detalhados sobre essa análise para as dimensões de qualidade abordadas neste estudo constam do Apêndice A - Tabelas ANOVA.

O teste estatístico de *Dunnett* foi aplicado após realização da ANOVA no intuito de verificar se as diferenças encontradas entre os grupos I a IV e o grupo controle eram significativas para cada dimensão de qualidade. Com isso, foi possível identificar mais claramente o impacto da adição do ruído sobre a acurácia.

Os resultados são expressos anexando o valor P (probabilidade de significância). Importante salientar que para todos os testes estatísticos realizados foi adotado um nível de significância de 0,05. A partir daí, foram consideradas diferenças estatisticamente significativas aquelas cujo $P<0,05$. Os resultados constam nas Tabelas 4.3, 4.4 e 4.5.

Figura 4.4: Diagramas de caixas para acurácia dos classificadores segundo o grupo experimental, enfocando DQ-ACC para cada atributo.



(a) Atributo Título.

(b) Atributo Resumo.

(c) Atributo Palavras-chave.

(d) Atributo Texto.

Fonte: Elaborado pelo autor.

Como se pode observar na Figura 4.4, os atributos ”Título” e ”Palavras-chave” (Figura 4.4a e Figura 4.4c) foram os mais afetados pela contaminação, sendo os demais pouco impactados pela introdução dos ruídos. Nota-se claramente que a acurácia variou entre 35% e 55% com tendência de queda, a qual reforça as diferenças entre os grupos I a IV e o controle; resultados confirmados na ANOVA e teste de Dunnett expostos na Tabela 4.3. Note que para os atributos ”Título” e ”Palavras-chave” as letras assinaladas em super-índice diferem, identificando mais claramente as diferenças estatisticamente significativas entre cada um dos grupos I a IV e o grupo C. Os três classificadores apresentam resultados próximos para todos os atributos, entretanto, cabe destacar que o efeito do ruído não prevalece para os atribu-

tos ”Resumo” e ”Texto”, nos quais a acurácia se manteve na faixa de 50% a 70% para os três classificadores. Os resultados da Tabela 4.3 permitem pressupor que não há diferenças estatisticamente significativas entre os classificadores, ou seja, comparativamente, os MVS, NBM e RL apresentaram a mesma acurácia em média para ”Título” (P=0,637), ”Resumo” (P=0,422) e ”Texto” (P=0,459) quando se analisou DQ-ACC. Exceto para o atributo ”Palavras-chave” em que os algoritmos classificadores apresentaram diferenças mais expressivas (P=0,043), nos demais a adição de ruído não prejudicou de forma significativa a acurácia dos algoritmos.

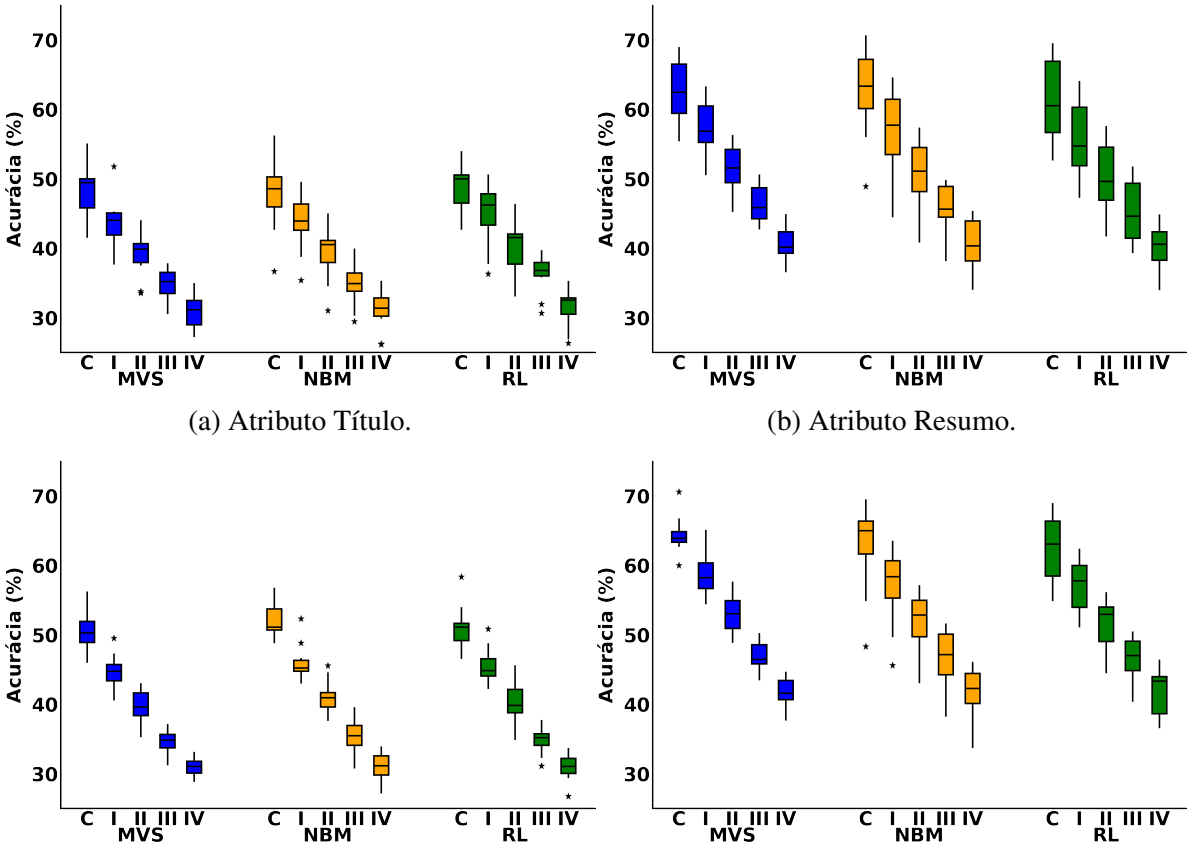
Tabela 4.3: Resultados para análise de DQ-ACC referente a acurácia média obtida para cada classificador por grupo, conforme o atributo.

		Grupo					
Atributo	Classificador	C	I	II	III	IV	Valor P**
Título	MVS	48,2[a]	46,1[b]	44,8[b]	42,9[b]	41,7[b]	0,637
	NBM	47,9[a]	46,8[b]	45,5[b]	43,8[b]	42,0[b]	
	RL	48,6[a]	48,2[b]	47,2[b]	46,2[b]	45,9[b]	
Resumo	MVS	62,8[a]	62,3[a]	61,8[a]	61,2[a]	60,5[a]	0,422
	NBM	62,5[a]	61,9[a]	61,4[a]	61,0[a]	61,0[a]	
	RL	61,2[a]	60,7[a]	60,3[a]	60,1[a]	60,8[a]	
Palavras-chave	MVS	50,4[a]	48,3[b]	46,2[b]	43,4[b]	41,9[b]	**0,043**
	NBM	52,2[a]	49,6[b]	47,3[b]	45,2[b]	43,8[b]	
	RL	51,1[a]	49,8[b]	48,5[b]	47,4[b]	46,6[b]	
Texto	MVS	64,4[a]	63,9[a]	63,2[a]	63,0[a]	62,2[a]	0,459
	NBM	62,6[a]	62,6[a]	62,4[a]	62,2[a]	62,3[a]	
	RL	62,6[a]	62,7[a]	62,5[a]	62,3[a]	62,2[a]	

[*] Letras distintas indicam diferença significativa entre os grupos pelo teste de Dunnett
[**] Valor p correspondente à comparação entre classificadores pela análise de variância - Two-Way ANOVA

Fonte: Elaborado pelo autor.

Figura 4.5: Diagramas de caixas para acurácia dos classificadores segundo o grupo experimental, enfocando DQ-COMP para cada atributo.



(a) Atributo Título.

(b) Atributo Resumo.

(c) Atributo Palavras-chave.

(d) Atributo Texto.

Fonte: Elaborado pelo autor.

A Figura 4.5 torna mais visível como a contaminação influenciou de forma significativa a acurácia obtida pelos classificadores. Dessa vez, porém, observa-se que todos os atributos foram afetados pelas condições experimentais, de modo que as diferenças entre os grupos são acentuadas e com tendência a queda na acurácia dos classificadores quando se comparam os grupos I a IV em relação ao grupo C. Para ”Titulo” e ”Palavras-chave” a acurácia variou entre 25% e 60%, enquanto que para ”Resumo” e ”Texto varia entre 35% e 70%. Os resultados da ANOVA e teste de Dunnett apresentados na Tabela 4.4 fornecem evidências mais seguras disso, uma vez que foi identificado que houve diferença estatisticamente significativa entre os grupos I a IV e o grupo C. Em relação aos classificadores, não foram verificadas diferenças significativas entre eles para nenhum dos atributos, sendo P>0,05 para todos. Tal resultado

evidencia que ao contaminar a BDC de modo a provocar a omissão de atributos e afetar a dimensão completude, houve um impacto significativo na acurácia média dos algoritmos classificadores. Cabe salientar que, assim como observado para a DQ-ACC, os atributos ”Título” e ”Palavras-chave” são os que apresentaram maior quantidade de pontos discrepantes, apesar de menor variabilidade na acurácia dentro de cada grupo.

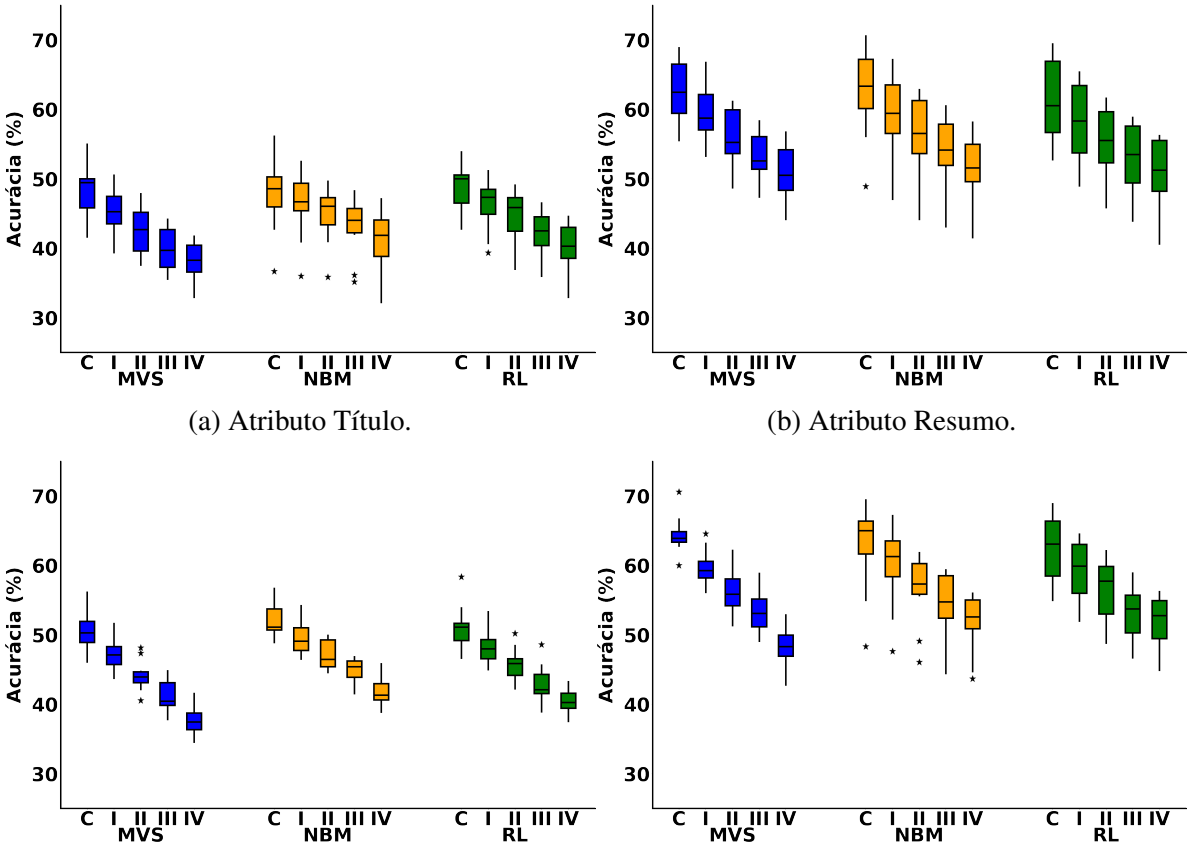
Tabela 4.4: Resultados para análise de DQ-COMP referente a acurácia média obtida para cada classificador por grupo conforme o atributo.

		Grupo					
Atributo	Classificador	C	I	II	III	IV	Valor P**
Título	MVS	48,2[a]	43,5[b]	39,0[b]	34,8[b]	31,0[b]	0,413
	NBM	47,9[a]	43,8[b]	39,5[b]	34,9[b]	30,9[b]	
	RL	48,6[a]	44,8[b]	39,5[b]	35,4[b]	30,9[b]	
Resumo	MVS	62,8[a]	57,6[b]	51,5[b]	46,4[b]	40,7[b]	0,332
	NBM	62,5[a]	56,6[b]	50,7[b]	45,6[b]	40,3[b]	
	RL	61,2[a]	55,6[b]	49,8[b]	45,3[b]	40,0[b]	
Palavras-chave	MVS	50,4[a]	44,8[b]	39,8[b]	34,6[b]	31,1[b]	0,076
	NBM	52,2[a]	46,1[b]	41,1[b]	35,6[b]	31,2[b]	
	RL	51,1[a]	45,6[b]	40,3[b]	34,9[b]	30,9[b]	
Texto	MVS	64,4[a]	58,8[b]	53,1[b]	46,9[b]	41,6[b]	0,282
	NBM	62,6[a]	56,9[b]	51,7[b]	46,7[b]	41,5[b]	
	RL	62,6[a]	57,0[b]	51,5[b]	46,4[b]	41,7[b]	

[*] Letras distintas indicam diferença significativa entre os grupos pelo teste de Dunnett
[**] Valor P correspondente à comparação entre classificadores pela análise de variância - Two-Way ANOVA

Fonte: Elaborado pelo autor.

Figura 4.6: Diagramas de caixas para acurácia dos classificadores segundo o grupo experimental, enfocando DQ-CONS para cada atributo.



(a) Atributo Título.

(b) Atributo Resumo.

(c) Atributo Palavras-chave.

(d) Atributo Texto.

Fonte: Elaborado pelo autor.

A Figura 4.6 também salienta que a acurácia dos classificadores foi impactada de forma considerável com a inserção dos ruídos que afetaram a consistência da BDC. Todos os atributos foram prejudicados pela inserção de ruído, com tendência à queda da acurácia quando se comparam os grupos I a IV em relação ao grupo C. Os resultados da Tabela 4.5 fornecem evidências de que houve diferença significativa entre os algoritmos para os atributos ”Título” e ”Palavras-chave” (P=0,041 e P<0,001). Especificamente em relação ao atributo ”Titulo”, o grupo I não diferiu de forma significativa do grupo C. O mesmo não ocorreu com os grupos II, III e IV em relação ao grupo C para os outros atributos. Diante de tais resultados, cabe salientar que a consistência foi afetada com o aumento da contaminação da BDC e tal fato ocorreu de forma mais drástica para ”Titulo” e ”Palavras-chave”, os quais a acurácia resultante da

classificação foi menor.

Tabela 4.5: Resultados para análise de DQ-CONS referente a acurácia média obtida para cada classificador por grupo conforme o atributo.

Atributo	Classificador	Grupo					Valor P**
		C	I	II	III	IV	
Título	MVS	48,2[a]	45,1[a]	42,6[b]	39,9[b]	38,0[b]	**0,041**
	NBM	47,9[a]	46,4[a]	44,9[b]	43,2[b]	41,3[b]	
	RL	48,6[a]	46,4[a]	44,1[b]	42,4[b]	40,1[b]	
Resumo	MVS	62,8[a]	59,6[b]	56,1[b]	53,3[b]	50,9[b]	0,622
	NBM	62,5[a]	59,0[b]	56,0[b]	53,7[b]	51,3[b]	
	RL	61,2[a]	58,1[b]	55,3[b]	52,9[b]	50,8[b]	
Palavras-chave	MVS	50,4[a]	47,2[b]	44,2[b]	41,3[b]	37,7[b]	**<0,001**
	NBM	52,2[a]	49,6[b]	47,1[b]	44,8[b]	41,9[b]	
	RL	51,1[a]	48,4[b]	45,8[b]	43,0[b]	40,4[b]	
Texto	MVS	64,4[a]	59,7[b]	56,3[b]	53,5[b]	48,4[b]	0,854
	NBM	62,6[a]	59,8[b]	56,5[b]	54,2[b]	51,7[b]	
	RL	62,6[a]	59,4[b]	56,6[b]	53,4[b]	51,8[b]	

[*] Letras distintas indicam diferença significativa entre os grupos pelo teste de Dunnett
[**] Valor P correspondente à comparação entre classificadores pela análise de variância - Two-Way ANOVA

Fonte: Elaborado pelo autor.

De maneira geral, os resultados apresentados apontam para a tendência de que o algoritmo NBM apresente mais pontos discrepantes na classificação em relação aos demais, o que indica maior variabilidade na acurácia dos *folds* durante o processo de classificação. Constata-se também que ”Título” e ”Palavras-chave” apresentaram menores valores de acurácia, além da tendência em introduzir mais pontos discrepantes e ressaltar diferenças entre os algoritmos, quando comparados aos demais atributos.

Neste capítulo foi possível realizar a comparação entre os algoritmos de classificação por meio de diferentes técnicas de validação. Foram testadas diferentes condições experimentais com o intuito de mensurar os reflexos da qualidade dos dados no processo de classificação. Constatou-se que a acurácia obtida pelos classificadores irá depender de alguns fatores, como: o método de contaminação proposto, o atributo avaliado e a existência de diferença significativa entre os algoritmos.

Capítulo 5

Conclusão

Este trabalho se propôs a comparar algoritmos de classificação, priorizando a avaliação de dimensões de qualidade de dados, no processo de classificação dos atributos textuais de uma base de dados científicos. Para isso, classificadores foram selecionados por meio da avaliação em uma base de teste e um experimento de simulação envolvendo diferentes condições experimentais foi delineado para avaliar o efeito da inserção de ruídos sobre a acurácia dos algoritmos de classificação. Dimensões de qualidade de dados quantitativas foram analisadas neste contexto, uma vez que constituem um aspecto relevante para identificação de possíveis forças e fraquezas nas bases de dados.

Diante dos resultados, foi possível comparar o desempenho dos algoritmos no processo de classificação de texto científico, analisar os efeitos provocados pelos distintos métodos de contaminação sobre as dimensões de qualidade de dados e identificar o impacto da adição de ruído no desempenho dos algoritmos de classificação, na medida em que se verificou que haviam diferenças estatisticamente significativas entre os grupos que definiram as condições experimentais.

Assim, os algoritmos selecionados (MVS, NBM e RL) na Seção 3.1 foram comparados sob diferentes métodos de validação e condições experimentais. Dito isso, apesar do algoritmo NBM ter apresentado bom desempenho e estabilidade na classificação, na BDC apresentou maior instabilidade nas classificações, apesar do menor tempo de execução alcançado.

Quanto ao MVS e RL, ambos obtiveram resultados semelhantes, com boa performance com dados de alta dimensionalidade, como é o caso de dados textuais. Apesar disso, os resultados na classificação do atributo ”Texto” sem a inserção de ruídos mostraram que o MVS, no geral, obteve resultados superiores e mais estáveis, mesmo com um conjunto de treino inferior aos demais algoritmos.

Os resultados das curvas ROC evidenciaram diferenças na classificação. O MVS foi o único algoritmo capaz de classificar de forma satisfatória artigos científicos provenientes da área de ”Ciências da Saúde”. Observou-se seu pior desempenho para artigos da área de ”Ciências Biológicas”. Já em relação ao NBM e RL, observou-se que ”Línguistica, Letras e Artes” foi a área em que os algoritmos apresentaram melhor desempenho no processo de classificação e ”Ciências Sociais Aplicadas” foram as de pior desempenho. Tal resultado revela que o desempenho dos algoritmos esteja associado ao tipo de texto envolvido na classificação.

Em relação às dimensões de qualidade de dados avaliadas (DQ-ACC, DQ-COMP e DQ-CONS), constatou-se que a DQ-COMP foi a mais afetada pela contaminação dos dados, seguida pela DQ-CONS e DQ-ACC, respectivamente. Na análise da DQ-COMP, a redução significativa da acurácia decorrente da contaminação pode ser explicada pela diminuição das informações disponíveis para a classificação. A DQ-CONS também foi afetada pela inserção dos ruídos de maneira contundente. Isso se deve à ambiguidade causada pelo método proposto para esta dimensão. Por fim, a DQ-ACC, considerando os atributos de maior dimensionalidade como ”Resumo” e ”Texto”, foi pouco afetada pelos métodos propostos, possivelmente devido à baixa probabilidade de contaminação dos caracteres e à metodologia adotada para o processo de classificação, os quais constituíram fatores que influenciaram nos resultados.

Por fim, ao avaliar o efeito da inserção de ruídos nos classificadores, foram observadas diferenças significativas entre os algoritmos para os atributos de menor dimensionalidade, isto é, para ”Título” e ”Palavras-chave”, possivelmente devido à maior instabilidade nas classificações. Logo, no caso de de textos científicos, estes atributos podem ser considerados fontes de dados menos apropriadas na avaliação do desempenho dos classificadores selecionados.

5.1 Problemas encontrados

Durante a elaboração do trabalho, foram identificadas algumas dificuldades, tais como:

- Encontrar trabalhos na literatura que abordassem de forma conjunta algoritmos de classificação e qualidade de dados, visto que não há uma relação direta entre qualidade de dados sob o ponto de vista das dimensões de qualidade e os tipos de ruídos considerados na área de mineração de dados, pois algumas dimensões de qualidade de dados podem ser avaliadas sob ruídos provenientes de dados com semânticas distintas;
- Elaborar métodos de contaminação que afetassem outras dimensões de qualidade de dados, devido às dificuldades em encontrar trabalhos na área que as definissem de maneira quantitativa e de mensurá-las numericamente a partir de sua definição.

5.2 Trabalhos futuros

Em relação à qualidade de dados, é possível explorar outros métodos para avaliar as dimensões de qualidade, bem como outras dimensões quantitativas. Quanto ao processo de classificação, há espaço para variação dos métodos utilizados, como no pré-processamento, na extração e na seleção de características, conforme mostra a Figura 2.1. Além disso, como apenas três algoritmos foram selecionados, torna-se interessante avaliar como outros classificadores, como os que estão presentes na Figura 2.2, influenciariam a qualidade dos dados sob o ponto de vista das dimensões de qualidade, podendo até ser estendido por meio de algoritmos baseados em aprendizado de máquina profundo, como as redes neurais artificiais.

Referências

AGGARWAL, C. C.; ZHAI, C. *A Survey of Text Classification Algorithms*. In: ____. *Mining Text Data*. Boston, MA: Springer US, 2012. p. 163–222. ISBN 978-1-4614-3223-4. Disponível em: <https://doi.org/10.1007/978-1-4614-3223-4_6>.

ALPAR, P.; WINKELSTRÄTER, S. *Assessment of data quality in accounting data with association rules*. *Expert Systems with Applications: An International Journal*, v. 41, p. 2259–2268, 2014.

BAHARUDIN, B.; LEE, L. H.; KHAN, K.; KHAN, A. *A Review of Machine Learning Algorithms for Text-Documents Classification*. *Journal of Advances in Information Technology*, v. 1, 2010.

BALLOU, D.; PAZER, H. *Modeling data and process quality in multi-input, multi-output information systems*. *Management science*, p. 150–162, 1985.

BATINI, C.; CAPPIELLO, C.; FRANCALANCI, C.; MAURINO, A. *Methodologies for Data Quality Assessment and Improvement*. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 41, n. 3, jul. 2009. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/1541880.1541883>.

BLAKE, R.; MANGIAMELI, P. *The Effects and Interactions of Data Quality and Problem Complexity on Data Mining.* p. 160–175, 2011.

BRASSEL, K.; BUCHER, F.; STEPHAN, E.-M.; VCKOVSKI, A. *CHAPTER FIVE* - completeness. In: GUPTILL, S. C.; MORRISON, J. L. (Ed.). *Elements of Spatial Data Quality*. Amsterdam: Pergamon, 1995, (International Cartographic Association). p. 81–108. ISBN 978-0-08-042432-3. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780080424323500124>.

BRINDHA, S.; PRABHA, K.; SUKUMARAN, S. *A survey on classification techniques for text mining*. p. 1–5, 2016.

CAI, L.; ZHU, Y. *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*. *Data Science Journal*, v. 14, 2015.

CHAWLA, N.; JAPKOWICZ, N.; KOŁCZ, A. *Editorial:* special issue on learning from imbalanced data sets. *SIGKDD Explorations*, v. 6, p. 1–6, 2004.

DALIANIS, H. *Evaluation Metrics and Evaluation*. In: ____. *Clinical Text Mining*. [S.l.]: Springer, Cham, 2018. p. 45–53. ISBN 978-3-319-78502-8.

FERREIRA, L. J. *Qualidade de dados:* requisitos para melhorar a confiabilidade da análise estatística. 2020.

FLECKENSTEIN, M.; FELLOWS, L. *Modern Data Strategy*. [S.l.]: Springer International Publishing, 2018. ISBN 978-3-319-68992-0.

GARCÍA, J. M. *Environment for the evaluation and certification of data products quality*. 2017.

GARCíA, S.; LUENGO, J.; HERRERA, F. *Data Preprocessing in Data Mining*. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 331910246X.

GE, M.; PERSIA, F. *A Survey of Multimedia Recommender Systems:* challenges and opportunities. *International Journal of Semantic Computing*, v. 11, p. 411–428, 2017.

GU, Q.; ZHU, L.; CAI, Z. *Evaluation Measures of the Classification Performance of Imbalanced Data Sets*. In: . [S.l.: s.n.], 2009. v. 51, p. 461–471. ISBN 978-3-642-04961-3.

GUALO, F.; RODRíGUEZ, M.; VERDUGO, J.; CABALLERO, I.; PIATTINI, M. *Data Quality Certification using ISO/IEC 25012:* industrial experiences. 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. Springer, 2009. Disponível em: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.

HEINRICH, B.; HOPF, M.; LOHNINGER, D.; SCHILLER, A.; SZUBARTOWICZ, M. *Data quality in recommender systems:* the impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets*, 2019.

HOO, Z. H.; CANDLISH, J.; TEARE, D. *What is an ROC curve? Emergency Medicine Journal*, *British Association for Accident and Emergency Medicine*, v. 34, n. 6, p. 357–359, 2017. ISSN 1472-0205. Disponível em: <https://emj.bmj.com/content/34/6/357>.

HOSKIN, T. *Parametric and Nonparametric:* demystifying the terms. 2012. Disponível em: <https://www.mayo.edu/research/documents/parametric-and-nonparametric-demystifying-the-terms/doc-20408960>.

HOSSIN, M.; SULAIMAN, M. N. *A Review on Evaluation Metrics for Data Classification Evaluations*. *International Journal of Data Mining & Knowledge Management Process*, v. 5, p. 01–11, 2015.

HOSSIN, M.; SULAIMAN, M. N.; MUSTAPHA, A.; WIRZA, R. *A Hybrid Evaluation Metric for Optimizing Classifier*. p. 165–170, 2011.

HUANG, J.; LING, C. *Using AUC and accuracy in evaluating learning algorithms*. *IEEE Transactions on Knowledge and Data Engineering*, v. 17, n. 3, p. 299–310, 2005.

ISO/IEC. *ISO/IEC 25040:2011 Systems and Software Engineering – Systems and Software Quality Requirements and Evaluation (SQuaRE) – Evaluation Process*. *ISO/IEC, International Standard*, 2011.

JONES, K. S. *A statistical interpretation of term specificity and its application in retrieval*. *J. Doc.*, 1972.

JURAFSKY, D.; MARTIN, H. J. *Speech and Language Processing*. [S.l.: s.n.], 2021.

KADHIM, A. I. *Term Weighting for Feature Extraction on Twitter:* a comparison between bm25 and tf-idf. In: *2019 International Conference on Advanced Science and Engineering (ICOASE)*. [S.l.: s.n.], 2019. p. 124–128.

KARR, A.; SANIL, A.; BANKS, D. *Data quality:* a statistical perspective. *Statistical Methodology*, v. 3, p. 6–7, 2003.

KOWSARI, K.; MEIMANDI, K. J.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D.; ID, L.; BARNES. *Text Classification Algorithms:* a survey. *Information (Switzerland)*, v. 10, 2019.

LANEY, D. *Infonomics: How To Monetize Manage and Measure Information As an Asset for Competitive Advantage*. [S.l.]: Routledge, 2017.

LI, L.; PENG, T.; KENNEDY, J. *Improving Data Quality in Data Warehousing Applications*. In: *ICEIS*. [S.l.: s.n.], 2010.

MAO, W.; MU, X.; ZHENG, Y.; YAN, G. *Leave-one-out cross-validation-based model selection for multi-input multi-output support vector machine*. *Neural Computing and Applications*, v. 24, 2012.

MCGILVRAY, D. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN 9780080558394.

MINAEE, S.; KALCHBRENNER, N.; CAMBRIA, E.; NIKZAD, N.; CHENAGHLU, M.; GAO, J. *Deep Learning–Based Text Classification:* a comprehensive review. *ACM Comput. Surv.*, *Association for Computing Machinery*, New York, NY, USA, v. 54, n. 3, abr. 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3439726>.

MOHAMMED, M.; KHAN, M.; BASHIER, E. *Machine Learning: Algorithms and Applications*. [S.l.: s.n.], 2016. 7 p. ISBN 9781498705387.

NARASIMHAN, H.; PAN, W.; KAR, P.; PROTOPAPAS, P.; RAMASWAMY, H. G. *Optimizing the Multiclass F-Measure via Biconcave Programming*. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. [S.l.: s.n.], 2016. p. 1101–1106.

RANAWANA, R.; PALADE, V. *Optimized Precision* - a new measure for classifier performance evaluation. In: . [S.l.: s.n.], 2006. p. 2254 – 2261.

REDMAN, T. C. *Data's Credibility Problem*. *Harvard Business Review*, 2013. Disponível em: <https://hbr.org/2013/12/datas-credibility-problem>.

REDMAN, T. C. *Bad Data Costs the U.S. $3 Trillion Per Year*. *Harvard Business Review*, 09 2016. Disponível em: <https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year>.

RODRÍGUEZ, M.; OVIEDO, J. R.; PIATTINI, M. *Evaluation of Software Product Functional Suitability:* a case study. *Software Quality Professional Magazine*, v. 18, 2016.

SAMMUT, C.; WEBB, G. I. *Encyclopedia of Machine Learning*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2011. ISBN 0387307680.

SCIELO. *Página inicial da SciELO*. 2021. <https://www.scielo.org/>. Acesso em: 08 de julho de 2021.

Seara Vieira, A.; BORRAJO, L.; IGLESIAS, E. *Improving the text classification using clustering and a novel HMM to reduce the dimensionality*. *Computer Methods and Programs in Biomedicine*, v. 136, p. 119–130, 2016. ISSN 0169-2607. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016926071530050X>.

SHAH, F. P.; PATEL, V. *A review on feature selection and feature extraction for text classification*. In: *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. [S.l.: s.n.], 2016. p. 2264–2268.

SRIVASTAVA, T. *Difference between Machine Learning & Statistical Modeling*. 2015. Disponível em: <https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/>.

SUBRAMANIAM, L.; ROY, S.; FARUQUIE, T.; NEGI, S. *A survey of types of text noise and techniques to handle noisy text*. In: . [S.l.: s.n.], 2009. p. 115–122.

THANGARAJ, M.; SIVAKAMI, M. *Text Classification Techniques:* a literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, v. 13, p. 117–135, 2018.

VASA, K. *Text Classification through Statistical and Machine Learning Methods:* a survey. In: . [S.l.: s.n.], 2016. v. 4, p. 655–658. ISSN 2321-9939.

VAUGHAN, T. *Multimidia: Making It Work*. [S.l.]: Osborne/McGraw-Hill, Berkeley, 1993.

VINCIARELLI, A. *Noisy text categorization*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 27, n. 12, p. 1882–1895, 2005.

WANG, R.; KON, H.; MADNICK, S. *Data Quality Requirements Analysis and Modeling*. In: . [S.l.: s.n.], 1993. p. 670 – 677. ISBN 0-8186-3570-3.

WONG, T.-T. *Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation*. *Pattern Recognition*, v. 48, 2015.

YANG, Y. *An Evaluation of Statistical Approaches to Text Categorization*. *Inf. Retr.*, *Kluwer Academic Publishers*, USA, v. 1, n. 1–2, p. 69–90, maio 1999. ISSN 1386-4564. Disponível em: <https://doi.org/10.1023/A:1009982220290>.

YING, X. *An Overview of Overfitting and its Solutions*. *Journal of Physics: Conference Series*, v. 1168, p. 022022, 2019.

ZHONGGUO, Y.; LI, H.; ALI, S.; AO, Y. *Choosing Classification Algorithms and Its Optimum Parameters based on Data Set Characteristics*. *Journal of Computers (Taiwan)*, v. 28, p. 26–38, 2017.

Apêndice A - Tabelas ANOVA

Tabela A-2: Análise de variância para a DQ-ACC considerando o método de contaminação de adição de caracteres.

Fonte da variação	SQ	GL	MQ	F	Valor p
Título					
Classificadores	16,68103	2	8,340514	0,453234	0,637
Grupos	737,9698	4	184,4925	10,02556	<0,001
Classificadores * Grupos	4,687695	8	0,585962	0,031842	0,01
Erro	2484,299	135	18,40221		
Total	3243,638	149			
Resumo					
Classificadores	57,39337	2	28,69669	0,868145	0,422
Grupos	55,90835	4	13,97709	0,422841	0,792
Classificadores * Grupos	5,533992	8	0,691749	0,020927	0,999
Erro	4462,452	135	33,05520		
Total	4581,287	149			
Palavras-chave					
Classificadores	64,28679	2	32,14339	3,224342	0,043
Grupos	1454,493	4	363,6232	36,47548	<0,001
Classificadores * Grupos	5,033374	8	0,629172	0,063113	0,01
Erro	1345,812	135	9,968978		
Total	2869,625	149			
Texto					
Classificadores	36,47090	2	18,23545	0,783186	0,459
Grupos	25,09732	4	6,274331	0,269474	0,897
Classificadores * Grupos	11,12230	8	1,390288	0,059711	0,999
Erro	3143,295	135	23,28367		
Total	3215,986	149			

Tabela A-2: Análise estatística pela ANOVA para a DQ-COMP considerando o método de contaminação de omissão de atributo.

Fonte da variação	SQ	GL	MQ	F	Valor p
Título					
Classificadores	24,80745	2	12,40372	0,890923	0,413
Grupos	5544,103	4	1386,026	99,55412	<0,001
Classificadores * Grupos	6,063745	8	0,757968	0,054443	0,01
Erro	1879,515	135	13,92233		
Total	7454,489	149			
Resumo					
Classificadores	49,37420	2	24,68710	1,112820	0,332
Grupos	8935,254	4	2233,814	100,6936	<0,001
Classificadores * Grupos	7,192469	8	0,899059	0,040527	0,01
Erro	2994,876	135	22,18427		
Total	11986,70	149			
Palavras-chave					
Classificadores	30,90037	2	15,45018	2,621615	0,076
Grupos	7761,694	4	1940,423	329,2546	<0,001
Classificadores * Grupos	8,659095	8	1,082387	0,183661	0,993
Erro	795,6068	135	5,893384		
Total	8596,860	149			
Texto					
Classificadores	40,46086	2	20,23043	1,278189	0,282
Grupos	8775,493	4	2193,873	138,6122	<0,001
Classificadores * Grupos	19,64551	8	2,455689	0,155154	0,996
Erro	2136,702	135	15,82742		
Total	10972,30	149			

Tabela A-3: Análise estatística pela ANOVA para a DQ-CONS considerando o método de contaminação de troca de atributo.

Fonte da variação	SQ	GL	MQ	F	Valor p
Título					
Classificadores	107,7453	2	53,87267	3,276447	0,041
Grupos	1369,628	4	342,4070	20,82462	<0,001
Classificadores * Grupos	45,43819	8	5,679774	0,345434	0,947
Erro	2219,725	135	16,44241		
Total	3742,536	149			
Resumo					
Classificadores	26,05831	2	13,02916	0,476832	0,622
Grupos	2367,514	4	591,8785	21,66116	<0,001
Classificadores * Grupos	9,347243	8	1,168405	0,042761	0,01
Erro	3688,795	135	27,32441		
Total	6091,714	149			
Palavras-chave					
Classificadores	226,0231	2	113,0115	19,99693	<0,001
Grupos	2320,524	4	580,1310	102,6518	<0,001
Classificadores * Grupos	21,30325	8	2,662906	0,471190	0,875
Erro	762,9450	135	5,651445		
Total	3330,796	149			
Texto					
Classificadores	6,025473	2	3,012737	0,157539	0,854
Grupos	2911,709	4	727,9272	38,06415	<0,001
Classificadores * Grupos	96,33276	8	12,04159	0,629669	0,752
Erro	2581,699	135	19,12370		
Total	5595,766	149			