Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Elaboração de Descritores para detecção de Risco de Conluio em Dados Públicos de Licitações de Obras

Roberta Costa Silva

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Flávio de Barros Vidal

Brasília
2021

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Elaboração de Descritores para detecção de Risco de Conluio em Dados Públicos de Licitações de Obras

Roberta Costa Silva

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Flávio de Barros Vidal (Orientador)
CIC/UnB

Prof. Dr. Maristela Terto de Holanda
CIC/UnB

Prof. Dr. Carla M. Chagas Cavalcanti Koike
CIC/UnB

Prof. Dr. João José Costa Gondim
Coordenador do Curso de Engenharia da Computação

Brasília, 05 de maio de 2021

Dedicatória

À minha família amada e ao querido São José.

Agradecimentos

Ao Professor Dr. Flávio de Barros Vidal pela orientação, incentivo e aprendizado proporcionado no desenvolvimento deste trabalho.

Ao Mestre e colega Marcos Cavalcanti Lima pela introdução ao problema e suporte ao longo do desenvolvimento deste projeto.

Ao colega Thiago Pinheiro pelo desenvolvimento do classificador *baseline* utilizado para comparação dos resultados deste trabalho.

À Universidade de Brasília, por proporcionar o desenvolvimento desta pesquisa.

Resumo

Neste trabalho, métodos automáticos de classificação envolvendo análises de grandes quantidades de dados são aplicados para a utilidade pública de detecção de conluio em contratos da gestão governamental. Esta análise se dá por meio da avaliação das bases textuais oficiais públicas, compostas pelos documentos contratuais divulgados publicamente, agregadas aos resultados de investigações oficiais para determinação do conjunto de características associadas à práticas de conluio. O descritor implementado por meio de técnicas de extração de características, baseado na frequência das palavras de um texto, em conjunto com o algoritmo de árvore de decisão foi capaz de classificar os documentos da base textual avaliando o risco de conluio associado. Aperfeiçoamentos futuros incluem analisar os resultados para a seleção de diferentes intervalos de frequência implementada como descritor.

Palavras-chave: Reconhecimento de Padrões, Classificação Textual, Extração de características, Conluio

Abstract

In this work, automatic classification methods involving the analysis of large amounts of data are applied for the public utility of collusion detection in government management contracts. This analysis takes place through the evaluation of the official public textual bases, composed of the contract documents disclosed, added to the results of official investigations to determine the set of associated characteristics. collusion practices. The descriptor implemented with characteristic extraction techniques based on the frequency of the words in a text, used with the decision tree algorithm was able to classify the documents of the textual basis evaluating the risk of associated collusion.

Keywords: Pattern Recognition, Textual Classification, Feature extraction, Collusion

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

Variáveis

μ	*Média*
ω	*Inclinação da reta*
σ	*Desvio padrão*
A	*Vetor de atributos*
b	*Valor de interceptação da reta no eixo y, bias*
C	*Vetor de classes*
D	*Partição de treinamento dos dados*
d	*Distância entre pontos no espaço dos atributos*
H	*Hipótese Bayesiana*
m	*Número de classes*
N	*Número de instâncias de dados, documentos*
n	*Número de atributos*
$P(A\|B)$	*Probabilidade condicional do evento A dado o evento B*
t	*Termo*
w	*Pesos*
X	*Vetor de valores dos atributos*
y	*Classe real ou predita*
Z	*Espaço de atributos*
z	*Coeficiente linear*

Capítulo 1

Introdução

Neste capítulo serão apresentados a motivação, a definição do problema de pesquisa, os objetivos e justificativa, para o desenvolvimento deste trabalho.

Ao final do capítulo, a organização do manuscrito é descrita, guiando a leitura deste trabalho.

1.1 Motivação

Este trabalho é motivado pela iniciativa de desenvolvimento de tecnologias utilizando algoritmos computacionais modernos para mineração de dados associadas ao combate à corrupção [10] [11] [12].

A tecnologia de mineração de dados implementada para o projeto é fundamentada na popularidade de processos de análise modernos e obtenção de informações de bases textuais, em contraste a técnicas tradicionais de recuperação de informação (IR), devido ao aumento expressivo de bases textuais com as mídias atuais, como *websites*, redes sociais e também por meio da digitalização de documentos [13].

O DOU, Diário Oficial da União, é um exemplo de nova base de dados textual disponível por meio digital <*https://www.gov.br/imprensanacional/pt-br*>. Com os dados de gestão governamental abertos para consulta digital são possíveis diversas análises para geração de conhecimento acerca dos processos conduzidos pelo governo, proporcionando maior transparência e integração entre a população e seus representantes.

Este trabalho propõe-se a analisar as bases textuais do DOU em busca de indícios da correlação entre publicações referentes a contratos públicos e o seu potencial risco de associação à prática de conluio.

1.2 Definição do Problema de Pesquisa

A pesquisa conduzida neste trabalho busca a elaboração de um descritor de características a ser utilizado junto ao processo de classificação para a avaliação de bases de texto públicas do Diário Oficial da União por meio de técnicas de mineração de dados textuais.

1.3 Objetivos

Nesta seção são expostos os objetivos gerais e específicos direcionadores deste trabalho.

1.3.1 Objetivos Gerais

O objetivo principal é o desenvolvimento de descritores da base textual para a classificação de publicações do diário oficial da União, avaliando o risco de conluio em contratos públicos.

1.3.2 Objetivos Específicos

Os objetivos específicos são:

1. Propor uma nova abordagem para a construção de descritores para bases textuais;
2. Analisar em situação real os descritores propostos;
3. Analisar os descritores propostos como entradas adicionais para modelos de redes neurais desenvolvidas para o projeto;
4. Avaliar os resultados obtidos dos descritores propostos com os demais existentes no estado-da-arte.

1.4 Justificativa

É notável a relação entre as estruturas implementadas no Brasil para contratações de empresas privadas de execução de obras públicas, como por exemplo obras de infraestrutura, e ações de corrupção [14].

O complexo modelo nacional de organização estrutural administrativa condiciona o processo para aquisição de serviços e produtos necessários à satisfação dos interesses públicos ao cumprimento de determinados regramentos, ora não tão eficazes para satisfazer os interesses da transparência, economia e legalidade dos atos públicos.

O meio de publicação das contrações públicas no Brasil é o Diário Oficial da União (DOU), que reúne todos os contratos públicos desde 1862, disponibilizando uma quantidade expressiva de informações para análise. Além das informações provenientes do DOU, processos de contratação governamentais estão associados aos documentos relacionados ao processo burocrático [15] [16].

Neste contexto, iniciativas utilizando tecnologias inovadoras são essenciais para, utilizando os dados públicos disponíveis, identificar, analisar e combater eventos de desvio de verba pública. Um exemplo de um projeto desenvolvido com esse objetivo é o projeto Serenata de Amor [10], o qual utiliza ciência de dados para analisar os gastos públicos e identificar comportamentos suspeitos na prestação de contas de entes públicos.

A investigação de ações de conluio pela Polícia Federal ao longo da história produziu estudos acumulados em uma base de conhecimento única que definem as metodologias [11] utilizadas pelos peritos profissionais. Este conhecimento estruturado para ser utilizado junto a ferramentas de inteligência é capaz de realizar a tarefa de identificação de processos fraudulentos [12]. A assimilação entre as informações disponíveis e métodos atuais de inteligência computacional se apresenta como uma ação efetiva e necessária para o combate à corrupção.

1.5 Organização do Manuscrito

Os capítulos deste trabalho foram organizados da seguinte forma: O Capítulo 2 Fundamentação Teórica, que consiste na apresentação cronológica dos métodos possíveis para a classificação textual e embasamento teórico para as discussões da metodologia proposta e dos resultados. O Capítulo 3 apresenta os Trabalhos Relacionados históricos na análise de técnicas de classificação textual e técnicas de extração de características. O Capítulo 4 apresenta a Metodologia Proposta para a implementação do trabalho. O Capítulo 5 compreende os Resultados obtidos com a implementação. Por fim, o Capítulo 6 apresenta a Conclusão e os Trabalhos Futuros pretendidos.

Capítulo 2

Fundamentação Teórica

Na primeira seção deste capítulo será explicado o que é classificação textual, a etapa de preparação dos dados antes da classificação e meios de avaliação da eficiência da classificação. A segunda seção descreverá os principais métodos para a classificação textual seguindo a ordem cronológica em que a abordagem passou a compor os meios para realização desta tarefa. Também realizando uma análise temporal, a terceira seção deste capítulo apresentará os métodos para a extração de características texto.

2.1 Classificação textual

O objetivo da tarefa de classificação é separar os dados em categorias de acordo com seus padrões. No contexto de dados textuais, as instâncias dos dados são documentos [17].

No processo de classificação textual, os documentos são categorizados de acordo com o assunto ou tópico, por exemplo, 'Política', 'História' ou 'Medicina', e é possível que mais de uma categoria sirva ao documento, designado um problema de múltiplas classes. Portanto, a classificação textual é interpretada como uma classificação binária (0, 1) de cada categoria para todos os documentos, determinando se cada documento é relacionado a um assunto ou não [17] [18].

A classificação é um problema de predição composto pela etapa de aprendizado, na qual as características de um conjunto de dados e suas importâncias são assimiladas pelo algoritmo para definir a categoria de uma instância do conjunto de dados [13].

A etapa de aprendizado pode ocorrer de forma supervisionada, em que há a análise de exemplos dos dados previamente classificados, ou de forma não-supervisionada, situação em que não estão disponíveis conjuntos de exemplos já categorizados.

Em uma representação formal, as características de uma instância são descritas pelo vetor X com o valor dos seus atributos [13].

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \tag{2.1}$$

Na etapa de aprendizado, é parametrizada a função f capaz de predizer a categoria de um exemplo representada pela classe y dado o seu vetor de características X [13]. A Subseção 2.2 apresenta os diferentes métodos para tal tarefa.

$$y = f(X) \tag{2.2}$$

É possível que um documento seja representado por *n-grams*, combinações de n palavras ou caracteres [17]. Em alguns casos, como na classificação de textos utilizando os métodos 2.2.5 e 2.2.1, é possível obter melhores performances e resultados utilizando a representação *em bi-grams* [19]. No caso da representação por palavras (*bag-of-words*) os atributos são cada termo do documento, neste caso x_n é uma palavra completa do texto.

Tratando-se de dados textuais, não é possível definir todos os atributos de uma classe considerando somente um documento, pois os dados são originários da linguagem natural, sejam textos escritos, discursos ou até mesmo conversas. Ou seja, os dados são formados das mais diversas formas por inúmeros termos da linguagem utilizada.

Por esse motivo, Bramer [17] define que, em tarefas utilizando dados textuais, os atributos serão todos os termos únicos do conjunto de documentos indicados pela posição em cada linha dos dados de aprendizagem. Assim, a representação de cada documento é dada por um vetor de atributos, que aponta a presença ou ausência de determinado atributo, indicando a quantidade em caso de repetição.

Como demonstrado no exemplo a seguir, os termos únicos de cada documento formam um dicionário em comum, definindo os atributos possíveis. Os textos dos documentos são representados por seus atributos presentes e quantidade dos termos do respectivo documento.

Documento 1: "A raposa e a lebre vivem na floresta"
Documento 2: "A raposa riu ontem de noite"
Documento 3: "A lebre ri hoje de dia"

Dicionário: {'a', 'raposa', 'e', 'lebre', 'vivem', 'na', 'floresta', 'riu', 'ri', 'ontem', 'hoje', 'de', 'noite', 'dia'}

Atributos do documento 1: {'2', '1', '1', '1', '1', '1', '1', '0', '0','0', '0', '0', '0', '0'}

Atributos do documento 2: {'1', '1', '0', '0', '0', '0', '0', '1', '0', '1', '0', '1', '1', '0'}

Atributos do documento 3: {'1', '0', '0', '1', '0', '0', '0', '0', '1', '0', '1', '1', '0', '1'}

É possível observar o aumento de dimensionalidade na representação dos dados, o que é apontado por Yan [20] como a maior dificuldade de trabalhar com dados textuais. Nota-se no exemplo o documento 3, mesmo possuindo somente 6 palavras, é representado por 14 atributos. Em textos maiores, essa característica é facilmente escalável, tornando-a ainda mais expressiva.

Dessa forma, é desejável e recomendado a utilização de técnicas de redução de dimensionalidade como extração e seleção de características para que os vetores de atributos sejam compostos idealmente por somente aqueles relevantes para a tarefa de classificação [21].

2.1.1 Pré-processamento de Dados Textuais

O pré-processamento dos dados textuais é fundamental como o primeiro método utilizado para redução de dimensionalidade de características.

Nesta etapa, um dos processos possíveis é a elaboração de uma lista de palavras comuns que podem ser retiradas do texto por não serem relevantes para a tarefa de classificação, chamadas de ***stop-words*** [17].

Para determiná-las, os termos são analisados de acordo com sua frequência e selecionados por meio de trabalho humano, considerando o conteúdo semântico dos termos conforme o contexto do *corpus* analisado [22].

Essa lista é variável até em uma mesma língua. Em português *stop-words* pode ser, por exemplo, artigos ('a(s)', 'o(s)', 'a(o)') ou ainda pronomes ('aquela(e)(s)') ou preposições ('em', 'para').

No exemplo da seção anterior:

Documento 1: "A raposa e a lebre vivem na floresta"

Documento 2: "A raposa riu ontem de noite"

Documento 3: "A lebre ri hoje de dia"

Dada a lista de *stop-words* = {'a', 'e', 'na', 'de'}, após retirá-las, os textos são:

Documento 1: "raposa lebre vivem floresta"

Documento 2: "raposa riu ontem noite"

Documento 3: "lebre ri hoje dia"

Ainda com o objetivo de reduzir a quantidade de palavras diferentes, a técnica de ***stemming*** consiste em indicar uma única representação de significado comum a palavras com variações de forma [17].

É possível observar nos documentos do exemplo a presença da desinência modo temporal nos termos 'ri' e 'riu'. Ou seja, são variações do verbo 'rir' para designação do tempo passado e presente. Os dois termos poderiam ser substituídos então por um único termo raiz, 'ri'.

Na etapa de pré-processamento ainda é possível retirar do texto as pontuações, símbolos, hífen, acentuação, ênclises de pronomes, valores numéricos, entre outros tipos de elementos de um texto. Deve ser analisado o conjunto de dados textuais e determinado o que pode ser útil para classificação e o que somente aumenta a dimensionalidade das características.

2.1.2 Avaliação do Modelo de Classificação

Um modelo, no contexto de classificação, é definição do comportamento uma função f (Equação 2.2), que compreende o padrão dos dados das possíveis categorias representadas pela variável dependente y, sendo assim capaz de determinar a classe a qual uma instância representada por X dos dados pertence [13]. Em um exemplo de aplicação, um modelo de classificação analisa o texto de postagens de redes sociais, representado por X, para classificá-lo na categoria $y =$ conteúdo ofensivo ou $y =$ conteúdo não ofensivo.

Para definir e testar a qualidade de um modelo, após o pré-processamento dos dados, o conjunto de todos os exemplos dos dados deve ser separado em conjunto de treinamento e conjunto de teste.

Nos algoritmos de classificação, o conjunto de treinamento é a entrada da etapa de aprendizado, na qual o modelo é parametrizado. O conjunto de testes é utilizado para validação da capacidade de predição do modelo em exemplos inéditos ao tempo de treinamento.

Matriz de confusão

Uma forma de avaliação do modelo utilizado é a matriz de confusão exemplificada de uma classificação binária na tabela abaixo, em que (C_1 ,C_2) = {'Positivo','Negativo'}. Os exemplos de teste são separados de acordo com o resultado da sua predição comparado com o resultado esperado, determinado pela classe real do exemplo.

Classes Reais	Resultado das predições	
Classe C_1	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Classe C_2	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Na diagonal de 'Verdadeiro', são contabilizados os exemplos que foram corretamente classificados de cada classe. Já na diagonal de 'Falso', são contabilizados os exemplos classificados incorretamente, isto é, um exemplo classificado como C_2 quando na realidade é da classe C_1 é um Falso Negativo, enquanto um exemplo classificado como C_1 quando na realidade é da classe C_2 é um Falso Positivo [17].

É necessário analisar a tolerância do erro do modelo quanto aos dois tipos de erros possíveis. Utilizando como exemplo a aplicação deste Projeto, em que o modelo está predizendo se um texto indica 'Alto risco de Fraude' (C_1) ou 'Baixo risco de fraude' (C_2), uma análise possível para os tipos de erro é: é desejável que mais textos sejam indicados como fraude para o trabalho de investigação posterior com o objetivo de diminuir a quantidade de fraudes despercebidas (Falsos Negativos), com a penalidade de potencialmente aumentar a quantidade de textos inocentes serem investigados (Falsos Positivos).

É adequado normalizar os valores de contagem de exemplos pela porcentagem da quantidade no conjunto completo de teste para melhor visualização de qual eventual classe o modelo está classificando de forma incorreta.

Métricas

As métricas comumente utilizadas para a avaliação de modelos de classificação textual são: acurácia, *precision*, *recall* e *f1-score*.

A acurácia (Equação 2.5) é proporção de exemplos corretamente classificados para ambas as classes no total de exemplos do conjunto de teste [17].

A medida *precision* (Equação 2.3) descreve a proporção de exemplos corretamente classificados na classe C_1 no total de exemplos que são realmente da classe C_1, total calculado pela soma dos exemplos classificados corretamente e incorretamente da classe C_1 ($VP + FP$) [17].

O *recall*, definido na Equação 2.4, refere-se a proporção de exemplos corretamente classificados na classe C_1 no total de exemplos do conjunto de teste, tanto os realmente da classe C_1 quanto os realmente da classe C_2 ($VP + VN$)[17].

A média harmônica das medidas *precision* e *recall* define a combinação para consideração das duas avaliações na métrica denominada *f1-score* (Equação 2.6).

$$Precision = \frac{VP}{VP + FP} \tag{2.3}$$

$$Recall = \frac{VP}{VP + VN} \tag{2.4}$$

$$\text{Acurácia} = \frac{TP + TN}{(VP + FP) + (VN + FN)} \tag{2.5}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.6}$$

O *f1-score* é calculado para cada classe do problema de classificação binária e o seu resultado final é média dos scores obtidos para as duas classes, chamado *f1-score macro*. Ele também pode ser calculado de forma balanceada *f1-score weighted*, em que seu valor final é definido por meio da média ponderada utilizado como peso as quantidades de instâncias de cada classe. Esse modo é indicado para avaliação da classificação em conjuntos desbalanceados, isto é, conjuntos que contêm mais exemplos de uma classe [23].

Curva Característica de Operação do Receptor (ROC) e AUC

A curva ROC (*Receiver Operating Characteristic*) é utilizada na avaliação de classificações binárias. A medida *Area Under the Curve* (AUC), quantifica a área abaixo da curva ROC.

A curva ROC é construída a partir da taxa de falsos positivos (TFP), abscissa, e da taxa de verdadeiros positivos (TVP), ordenada.

A TFP mede a proporção de exemplos da classe C_2 (negativos) incorretamente preditos C_1 (positivos) na quantidade total de exemplos da classe C_2, fração descrita na Equação 2.8 pela quantidade de exemplos falsos positivos (FP) dividido pela quantidade total de exemplos negativos [24].

A TVP mede a proporção de exemplos corretamente preditos da classe C_1 no total de exemplos da classe C_1, fração descrita na Equação 2.7 pela quantidade de exemplos verdadeiros positivos (VP) dividido pela quantidade de exemplos positivos [24].

$$TVP = \frac{VP}{VP + FN} \tag{2.7}$$

$$TFP = \frac{FP}{FP + VN} \tag{2.8}$$

O modelo ideal, que não comete nenhum erro nas predições, possui AUC = 1. No ponto (0,1), as predições do modelo foram totalmente corretas para a classe C_1 e não houve nenhum erro nas predições da C_2, sendo possível afirmar que o modelo predisse todos os exemplos de teste corretamente. Nesse caso a AUC é igual a 1. Já no ponto (1,1), as predições do modelo foram totalmente corretas para a classe C_1 (positivos) e totalmente incorretas para a classe C_2 (negativos). Nesta ocasião, o modelo predisse a mesma classe para todos os exemplos. Portanto, é desejável que a curva ROC apresente-se no formato semelhante aos formatos da curvas verde e laranja da Figura 2.1 [24].

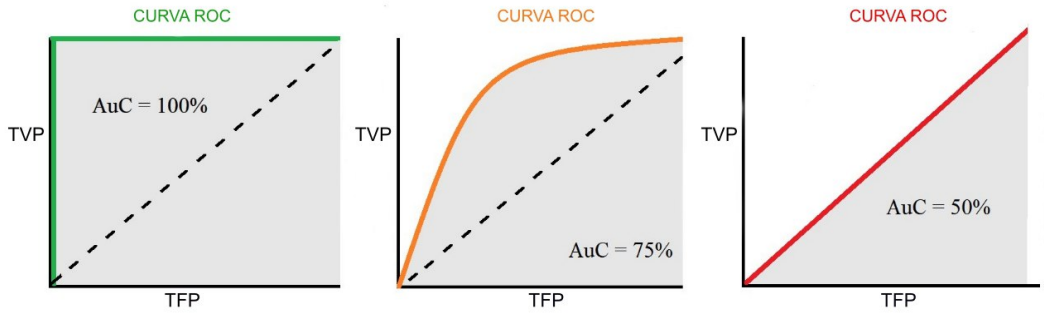

Figura 2.1: Exemplos de curvas ROC com comportamento ideal, desejável e indesejado, da esquerda para direita [1].

2.2 Métodos de Classificação Textual

Esta seção se dedicará a exposição dos métodos estatísticos e de aprendizado de máquinas para classificação de textos em ordem cronológica.

2.2.1 Modelos Bayesianos

O modelo estatístico *naive Bayes* foi criado a partir do Teorema (Equação 2.9) de Thomas Bayes [13], que descreve a probabilidade condicional $P(A|B)$ de um evento A ocorrer posteriormente à ocorrência de um evento B. Aplicando ao contexto de decisão de classificação, $\mathbf{X}$ contém os atributos de uma instância de dados e $\mathbf{H}$ é uma hipótese da forma: X pertence a determinada classe C [13].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2.9}$$

$P(H|X)$ é a probabilidade da hipótese H de X pertencer à classe C, dado que os atributos de X são conhecidos. De modo análogo, $P(X|H)$ é a probabilidade dos valores esperados dos atributos em X serem satisfeitos quando já foi determinado que X pertence à classe C. As probabilidades independentes a eventos anteriores, $P(X)$ e $P(H)$, correspondem à probabilidade de uma instância dos dados apresentar os valores dos atributos observados em X e à probabilidade de uma instância dos dados pertencer à classe C, respectivamente [13].

Um classificador naive Bayes define a probabilidade de um conjunto de atributos pertencer a uma determinada classe. O classificador calcula e compara a probabilidade do conjunto de atributos X descrever todas as classes (C_1,C_2,...,C_m) presentes e decide pela mais provável, de acordo com a Equação 2.10 [13].

$$P(C_i|X) > P(C_i|X) \quad para \quad 1 \leq j \leq m, j \neq i \tag{2.10}$$

De acordo com o Teorema de Bayes, $P(C_i|X)$ é calculado através da Equação 2.11:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}, \tag{2.11}$$

onde $P(X)$ é uma constante para todas as classes e $P(C_i)$ pode ser estimada a partir da divisão da quantidade de exemplos da classe C_i no conjunto de dados pela quantidade total de exemplos no conjunto de dados [13].

Com objetivo de reduzir o custo computacional para o cálculo da probabilidade $P(X|C_i)$, classificadores desse tipo assumem que, em uma determinada classe, as características são independentes entre si [13]. Isto é, a ocorrência de um atributo A_i com o valor x_i não interfere na probabilidade de ocorrência de outro atributo A_j com um valor x_j. A simplificação determina o nome do modelo *naive* [13], ingênuo ou simples em português.

Essa hipótese, chamada independência condicional de classe, é demostrada na Equação 2.12, em que as probabilidades dos valores referente aos atributos são calculadas individualmente [13].

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times P(x_3|C_i) \times ... \times P(x_n|C_i) \tag{2.12}$$

Desse modo, a probabilidade de uma instância X dos dados apresentar determinados valores de atributos dado que pertence à classe C_i será o produtório das probabilidades de ocorrência cada atributo na classe C_i [13].

Há duas formas de calcular $P(x_k|C_i)$. Na ocasião dos atributos serem categóricos, $P(x_k|C_i)$ é o número de exemplos do conjunto de dados que pertencem à categoria C_i dividido pelo número total de exemplos da classe C_i no conjunto de dados [13].

Destaca-se a capacidade de modelos Bayesianos lidarem com atributos de forma categórica.

Quando os atributos são valores contínuos, assume-se que os dados possuem uma distribuição normal(Figura 2.2) com uma média μ e desvio padrão σ, definida pela Equação 2.13 [13].

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\mu^2}} \tag{2.13}$$

$$P(x_k|C_i) = g(x, \mu, \sigma) \tag{2.14}$$

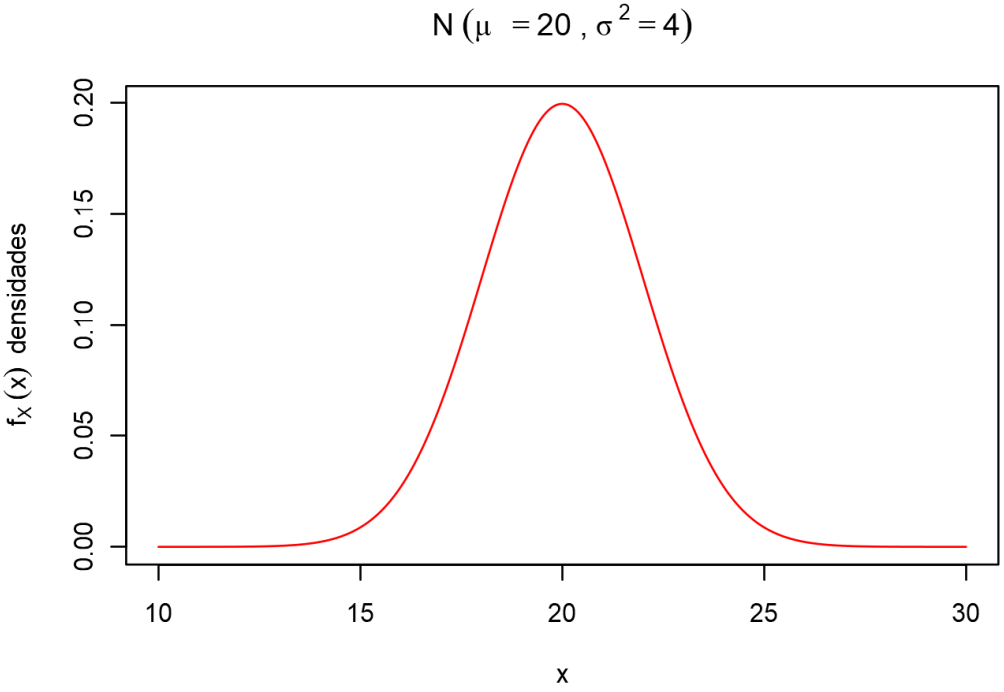

Figura 2.2: Distribuição normal: Também chamada de distribuição gaussiana, a distribuição normal relaciona a frequência (na imagem, representada por $f_x(x)$densidade) dos dados (x) utilizando os parâmetros de média (μ) e desvio padrão (σ), também denominado grau de dispersão [2].

2.2.2 *Nearest neighbor*

Cover [25] no estudo de aplicação do método de vizinho mais próximo para a classificação, cita o trabalho de 1951 de Fix e Hodges [26], que define a regra de k_n-*nearest neighbor*. Essa regra define que uma classe de um ponto pode ser representada pelos k vizinhos desse ponto, isto é, pelos pontos com determinado nível de proximidade.

Segundo Han [13], a classificação com este tipo de abordagem utiliza a similaridade entre um conjunto de n atributos determinado por uma tupla, que representa as coordenadas dos pontos mencionados em um espaço de n dimensões.

A proximidade entre os pontos de acordo com Han[13] é descrita pela distância euclidiana (Equação 2.15) entre as tuplas $X_1 = (x_{11}, x_{12}, ..., x_{1n})$ e $X_2 = (x_{21}, x_{22}, ..., x_{2n})$.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{2.15}$$

O modelo permite a variação da quantidade k de pontos vizinhos, que devem ter suas distâncias combinadas para a avaliação da menor distância do exemplo sendo classificado, indicado pelo ponto de interrogação na Figura 2.3. Após a determinar os vizinhos mais próximos, a classe predita para o exemplo é a mais frequente dentre as classes dos k vizinhos [17].

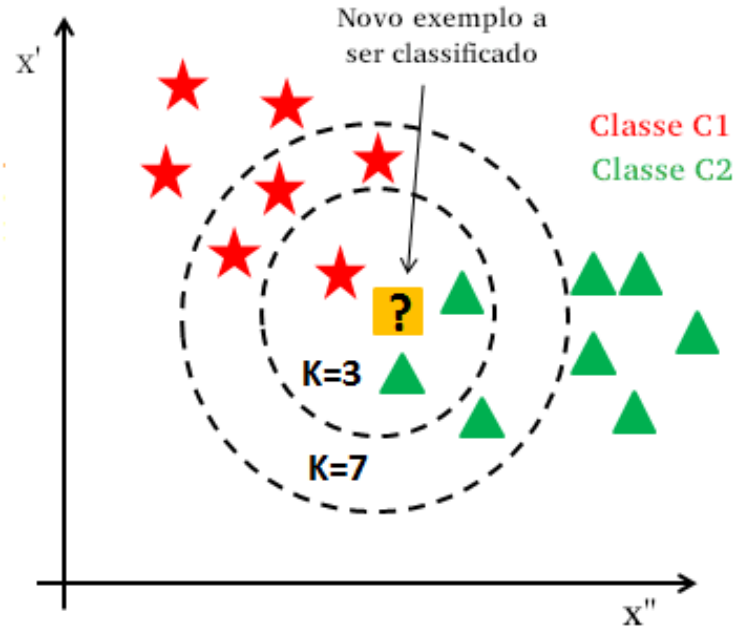

Figura 2.3: k-NN: Na imagem, os atributos x' e x'' descrevem os exemplos de treinamento, estrelas vermelhas e triângulos verdes, no espaço de atributos. $k = 3$ e $k = 7$ referem-se à quantidade de vizinhos [3].

Existem outras formas de cálculo de distância entre pontos, como a distância *Manhattan* [17] e a distância *Levenshtein* [27].

Para coerência na análise das distâncias é recomendada a normalização dos valores dos atributos, sendo geralmente utiliza a denominada max-min (Equação 2.16), em que o valores são transformados para o intervalo [0,1], sendo min_A o menor valor observado para o atributo A e max_A o maior valor [13].

$$\text{valor}_{[0,1]} = \frac{valor - min_A}{max_A - min_A} \tag{2.16}$$

Especificamente na tarefa de classificação textual, Liu [28] descreve que, para testar a classificação de documento, a similaridade é calculada para cada documento vizinho como um *score*, que define os pesos de cada classe presente nos vizinhos.

O resultado da média ponderada dos pesos dos vizinhos de cada classe é utilizada para a comparação com o documento sendo classificado. Assim, os limites de similaridade de cada classe podem ser definidos para fornecem a informação da classe com exemplos mais similares ao documento em questão.

2.2.3 Árvores de Decisão

A classificação por meio de árvores de decisão é realizada por meio da construção de uma estrutura em formato de árvore, observada na Figura 2.4, composta por [13]:

- Nós de decisão (internos): representam o teste de cada atributo de acordo com as classes. Seja X o vetor de atributos, cada elemento é testado seguindo as regras de separação definidas para os nós internos;
- Galhos: são os possíveis resultados para o teste dos nós. A decisão em cada nó interno forma o caminho até a classe determinada nos nós folha;
- Nós folha: indicam a categoria da instância do dado, resultado da comparação com as regras dos nós internos.

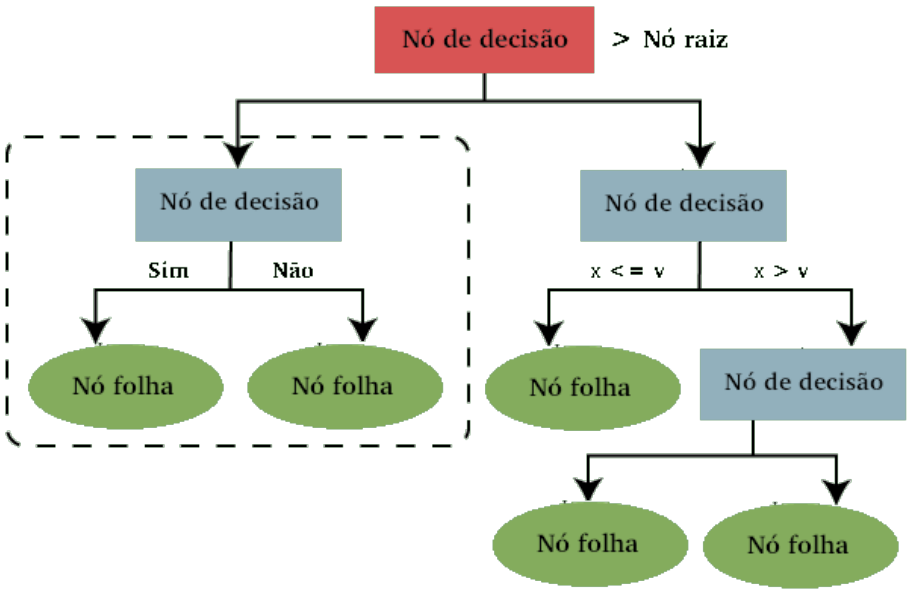

Figura 2.4: Árvore de decisão: Estrutura de uma árvore de decisão, com uma sub-árvore em destaque, com galhos para atributos categóricos ('sim'/'não') e atributos numéricos (v) [4].

Podem ser utilizadas características categóricas ou numéricas [13]. Por exemplo, na classificação de pessoas vítimas de Covid-19, sendo $\{0, 1\}$ as possíveis classes, representando 'recuperação' (0) e 'óbito' (1), a característica 'fumante' pode utilizada para descrever a amostra da população analisada. Nesse caso, são possíveis respostas 'sim' e 'não'. Seguindo o exemplo, uma característica numérica possível é a idade das pessoas.

As regras de separação formam o conjunto de regras a serem testadas para realizar a classificação de uma instância. No caso de um atributo categórico, os galhos de separação assumem os próprios valores do atributo (por exemplo, {'sim','não'}). E, quando há atributo numérico, um valor contínuo é definido para a separação ocorrer seguindo as condições $valor\_atributo \leq valor\_separação$ e $valor\_atributo > valor\_separação$ [13].

A maioria dos algoritmos de árvore de decisão constrói a estrutura de decisão de forma recursiva, começando a separação dos dados pelo nó raiz (indicado na Figura 2.4), percorrendo abaixo as decisões seguintes, em uma estratégia de "dividir para conquistar" [13].

Em ordem cronológica, CART [29], ID3 [30] e C4.5 [31] são algoritmos de árvore de decisão popularmente utilizados para comparações de performance e resultado entre diferentes modelos [13].

Inicialmente, o modelo calcula o quão bem cada característica descreve as classes, utilizando as medidas como **índice gini**, ***information gain*** (Subseção 2.3.1) ou *gain ratio*.

A implementação CART utiliza o índice gini como métrica da qualidade da separação dos dados. Já a implementação ID3 utiliza o *information gain* e, sua implementação sucessora, C4.5, utiliza o *gain ratio*.

São definidos os critérios de separação (galhos) de cada nó e uma característica que não separa os dados nas classe com 100% de acertos é considerada impura. O objetivo é ter a separação dos dados com menor índice de impureza [13].

Voltando ao exemplo do estudo de vítimas do Covid-19, a separação das pessoas pelo atributo fumantes seria pura se todos os não fumantes fizessem parte da classe de 'recuperação' e todos os fumantes fizessem parte da classe de 'óbitos'. Dessa forma, o atributo descreveria da melhor forma possível uma pessoa no universo dos dados e sua predição de classe.

O **índice de gini** é calculado de acordo com a Equação 2.17, em que D é a partição de treinamento dos dados, m é a quantidade de classes possíveis e p_i é a probabilidade dos atributos da partição D serem da classe de referência do índice i [13].

$$Gini(D) = 1 - \sum_{i=1}^{m} (p_i)^2 \tag{2.17}$$

$$Gini_{min} = 1 - (1^2) = 0 \tag{2.18}$$

$$Gini_{max} = 1 - (0,5^2 + 0,5^2) = 0,5 \tag{2.19}$$

Sendo $A = \{a_1, a_2\}$ o conjunto de atributos que descreve os dados e a_1 um atributo categórico {'sim','não'}. O índice de gini para a partição D, referente somente ao atributo a_1, é calculado para cada valor possível de a_1:

$Gini(D_1)$ = 1 - (probabilidade de uma amostra do conjunto de dados com a_1 = 'sim' ser da classe $C_1)^2$ - (probabilidade de uma amostra do conjunto de dados com a_1 = 'sim' ser da classe $C_2)^2$ e

$Gini(D_2)$ = 1 - (probabilidade de uma amostra do conjunto de dados com a_1 = 'não' ser da classe $C_1)^2$ - (probabilidade de uma amostra do conjunto de dados com a_1 = 'não' ser da classe $C_2)^2$.

O índice de gini resultante de D referente ao atributo a_1 é a média ponderada (Equação 2.20) de $Gini(D_1)$ e $Gini(D_2)$.

$$Gini_{a_1}(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \tag{2.20}$$

A cada nó de decisão, representado por um atributo dos dados, a classificação da instância deve ser aprimorada, isto é, o índice de impureza das separações de cada nó seguinte deve ser menor que o índice do nó anterior. Caso essa condição não seja atendida, o nó se torna uma folha, indicando a categoria em que a instância foi classificada [13].

Uma vantagem da utilização de árvores de decisão para classificação é a simplificação da representação das decisões em formato de árvore, que é facilmente assimilada por humanos. Além disso, árvores de decisão possuem capacidade para lidar com dados faltantes (*missing data*) e com dados de alta dimensionalidade, característica desejável na classificação textual [13].

2.2.4 Modelos Lineares

Em uma tarefa de classificação binária, o conjunto de dados pode apresentar uma relação linear. Isso significa que é possível separar os componentes de cada categoria por uma reta, no caso dos dados serem descritos por dois atributos, ou por um hiperplano [13].

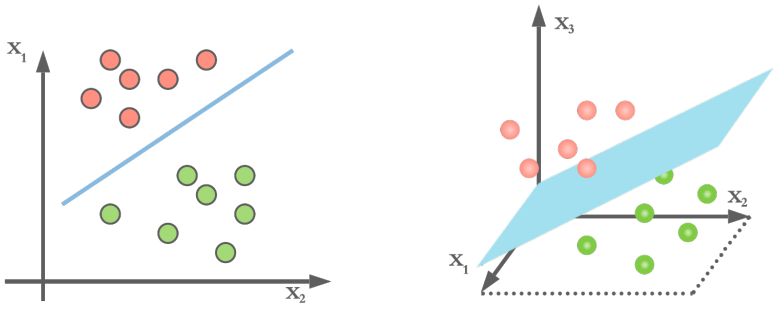

Figura 2.5: Dimensões (n) dos atributos (X_n), se $n = 2$, o conjunto de dados é divido por uma reta, se $n = 3$, o conjunto de dados é divido por um plano.

Na Figura 2.5, X_1, X_2 e X_3 são os vetores de atributos dos dados na forma descrita pela Equação 2.1.

Sendo n a quantidade de dimensões do espaço sendo analisado, definido pela quantidade de atributos que descrevem os dados, o hiperplano é um subespaço vetorial com n-1 dimensões. O conceito é utilizado para generalizar a forma geométrica na ocasião dos dados serem definidos por mais atributos. Quando três atributos, como na Figura 2.5, o hiperplano é um plano comum.

Modelos lineares são indicados para classificação de textos [32], pois possuem capacidade para lidar com dados representados por espaços de alta dimensionalidade [21], característica de dados textuais.

Regressão Logística

A regressão é utilizada para predição numérica [13], diferente da tarefa de classificação, que realiza a predição da classe.

Um método comum de regressão é a linear. Sendo **y** o valor numérico resultado da predição para determinado valor de atributo (**x**), **b** o valor em que a reta intercepta o eixo y e ω a inclinação da reta, o modelo linear é descrito pela função [13]:

$$y = b + \omega x \tag{2.21}$$

Dessa forma, a predição é determinada pela reta que descreve os dados da melhor forma possível, isso significa, a reta que produz predições com o menor erro para dados reais [13].

Ainda, considerando **b** e ω pesos genéricos w_0 e w_1, é possível expressar a função como [13]:

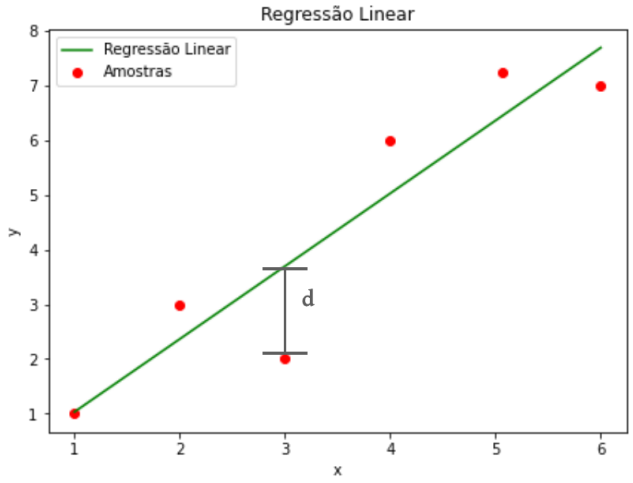

Figura 2.6: Regressão Linear: A inclinação da reta de predição é determinada por $\omega = (\Delta y_i)/(\Delta x_i)$ [5].

$$y = w_0 + w_1 x \tag{2.22}$$

O problema de predição utilizando a regressão linear é resolvido com a determinação dos valores dos pesos utilizando o método dos quadrados mínimos, encontrando a reta que minimiza a soma das distâncias **d** de todas as amostras e a reta [13].

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{|D|}(x_i - \overline{x})^2} \tag{2.23}$$

$$w_0 = \overline{y} - w_1 \overline{x} \tag{2.24}$$

Na equação 2.23, o peso w_1 é calculado a partir das distâncias entre os pontos (x_i, y_i), que representam os dados reais, e a média ($\overline{x}$ e $\overline{y}$) de todos os valores observados em x e y. O peso w_0 é definido na equação 2.24 pela média dos valores em y e x e o peso w_1.

A **Regressão Logística**, é um método utilizado para a tarefa de classificação descrito pela função logística, definida pela equação 2.25.

Nessa função, z é uma soma linear do parâmetro α e os produtos entre os atributos X_n e parâmetros β_n, que definem a classificação, como os pesos na regressão linear[33].

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2.25}$$

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{2.26}$$

A Figura 2.7 mostra os dados separados em y = 1 e y = 0 de acordo com a função logística. Nota-se que, diferente da regressão linear em que os valores preditos eram contínuos, no método de regressão logística o resultado é discreto e pertence ao intervalo [0,1], devido à função logística descrever uma probabilidade [33].

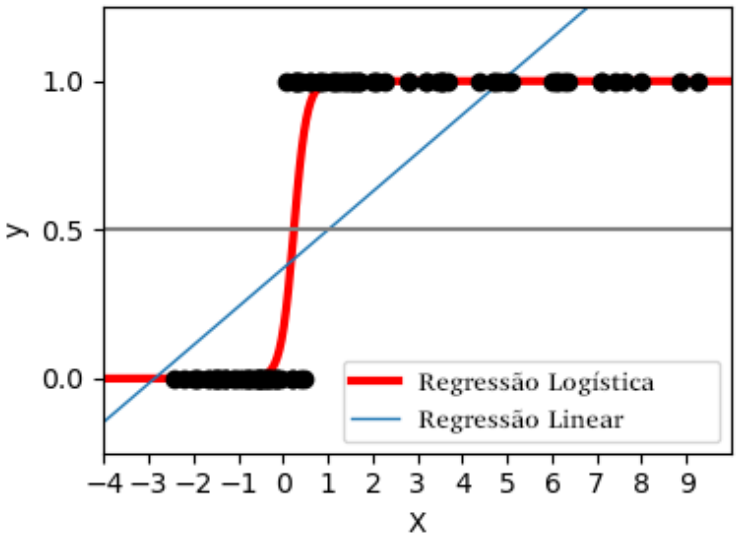

Figura 2.7: Regressão logística: A função logística descreve a probabilidade **y** entre 0 e 1 dado determinado **x** [6].

A regressão logística é utilizada para classificação de textos. Fan [32] descreve a biblioteca LIBLINEAR com suporte a regressão logística para problemas de classificação com dados em larga escala, caso da classificação textual.

Os autores do artigo demonstram que a implementação de regressão logística da biblioteca, embora apresente maior tempo de treinamento, atinge resultado de acurácia similares aos de modelos SVM, que possuem implementação mais complexa.

2.2.5 *Support Vector Machine* (SVM)

Segundo Cortes e Vapnik [34], criadores deste modelo de classificação, dados vetores de entrada das características, como X (equação 2.1), projetados em um espaço de atributos Z com n dimensões, sendo n a quantidade de atributos, o método SVM consiste em encontrar o melhor hiperplano para divisão dos vetores de entrada nesse espaço.

O método SVM consiste em, além de determinar o hiperplano de separação, definir as maiores margens possíveis (*maximum marginal hyperplane*) de divisão entre os dados [13].

Os *Support Vectors* são os vetores correspondentes a um pequeno subgrupo dos dados de treinamento, utilizados para delimitação das margens de separação dos dados [34], conforme apontado na Figura 2.8.

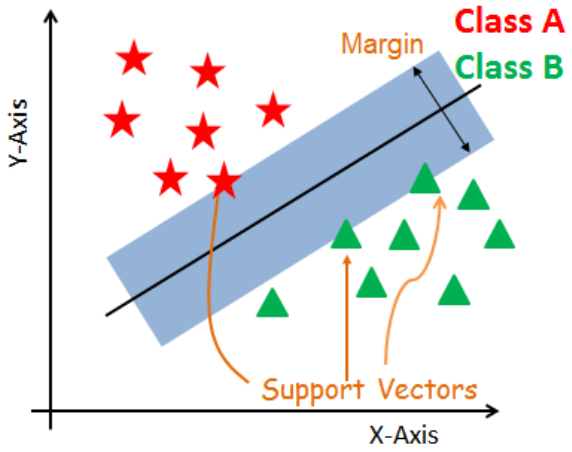

Figura 2.8: Quando Z possui duas dimensões, o hiperplano de divisão dos dados é uma reta [7].

Cortes [34] define que, sendo o hiperplano de separação uma reta definido pela na equação 2.27, os pesos w_0 são definidos por uma combinação linear (equação 2.28) entre os *support vectors*.

A superfície de separação é definida após o mapeamento do vetor de entrada no vetor de características definido no espaço de atributos. Este mapeamento é feito de acordo com um conjunto funções matemáticas, denominado *kernel*, podendo ser lineares, polinomiais ou funções de base radial (RBF), de acordo com a dispersão dos dados no espaço. Posteriormente ao mapeamento, um separador linear é determinado a partir da definição dos pesos e de b, denominado o viés, componente somada para permitir a movimentação do hiperplano de separação acima ou abaixo da coordenada de origem do espaço de atributos.

A equação linear 2.29 é a função de decisão $I(x)$, em que x_i são os *supported vectors* e x é o vetor X no espaço de atributos. A função de decisão é responsável por classificar os vetores de entrada desconhecidos.

$$0 = b_0 + w_0 x \tag{2.27}$$

$$w_0 = \sum_{\text{support vectors}} \alpha_i x_i \tag{2.28}$$

$$I(x) = sign(\sum_{\text{support vectors}} \alpha_i x_i \cdot x + b_0) \tag{2.29}$$

Quando, a princípio, os dados não são linearmente separáveis, são utilizados *kernels* polinomais ou RBF.

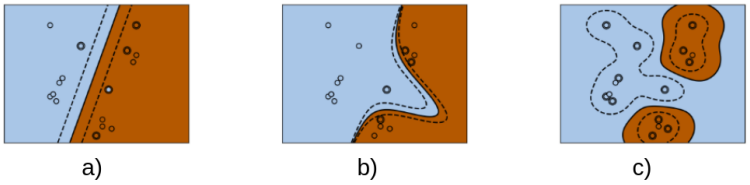

Figura 2.9: Apresentação dos hiperplanos de separação dos dados criados a partir de *kernel* a) linear, b) polinomial e c) RBF [8]

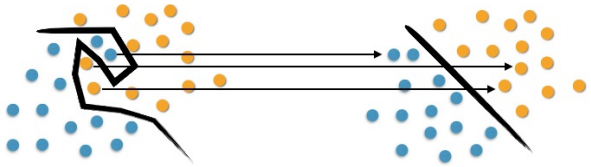

Figura 2.10: Representação do mapeamento dos dados realizado por meio de kernels não lineares para uma expressão linear [9]

Os dados X de um conjunto classificados de acordo com as categorias $y_n \in -1, 1$ são considerados linearmente separáveis se existir um vetor de pesos w e determinado b escalar que satisfazem [34]:

$$\begin{aligned} w \cdot x_i + b \geq 1 \quad se \quad y_i = 1 \\ w \cdot x_i + b \leq -1 \quad se \quad y_i = -1 \end{aligned} \tag{2.30}$$

O melhor hiperplano de separação, o hiperplano ótimo, é definido pela maior margem de separação entre os vetores de dados. Outro parâmetro que propicia um melhor resultado geral é utilizar a menor quantidade de *support vectors* mantendo a qualidade do modelo, pois assim o modelo possuirá maior capacidade de generalização, isto é, cometerá menos erros no conjunto de dados de teste [34].

Joachims [18], descreve por meio das propriedades de dados textuais evidências capazes de suportar a afirmação que modelos *Support Vector Machine* (SVM) são adequados para a classificação textual.

Essas propriedades são:

- Grande quantidade de características: Conforme exposto na seção 2.1, dados textuais possuem grande quantidade de atributos, definidos pelos termos do texto, o que produz espaços de alta dimensionalidade;
- Poucos termos irrelevantes em comparação a quantidade total de características: Mesmo utilizando técnicas de Extração e Seleção de características (Seção 2.3) para redução de dimensionalidade, ainda muitos termos devem ser utilizados para a classificação;
- Os vetores de representação dos documentos possuem alta dispersão: Como demonstrado no exemplo da seção 2.1, a representação dos documentos é feita a partir da indicação de presença ou não de um atributo dentre todos os atributos únicos de todos os documentos, produzindo vetores populados majoritariamente por zeros;
- Geralmente dados textuais são separáveis linearmente: O modelo SVM é utilizado para determinar a separação linear dos dados.

2.3 Extração e Seleção de Características

John [35] propõe dois tipos de métodos para determinar um subconjunto de características que será utilizado pelo modelo de forma a selecionar, idealmente, somente aquelas que trarão a melhor performance.

Para tal, é definido o conceito de relevância em dois graus: fraca e forte. As características que possuem a possibilidade de aumentar a acurácia da predição são as de relevância fraca, enquanto as que não podem ser retiradas sem perda de acurácia são as de relevância forte. As que não são classificadas fracas ou fortes são irrelevantes, pois não aumentam a acurácia da predição e podem ser retiradas.

O tipo de seleção de características proposto como melhor pelos autores do artigo citado é chamado *Wrapper*, em que a acurácia do modelo de classificação é utilizada como a função de avaliação de qualidade das características.

Métodos de seleção de características do tipo *Wrapper* não são adequados à classificação textual, pois, conforme apresentado nas seções anteriores, dados textuais possuem alta dimensionalidade de atributos, o que torna oneroso ser necessário repetir a classificação para avaliar os atributos [36].

O segundo tipo é o *Filter*, descrevendo métodos de seleção que funcionam de forma independente ao algoritmo de classificação. Estes são comumente utilizados para classificação de texto [36]. Nesse tipo, os atributos considerados irrelevantes são retirados do conjunto de dados anteriormente à realização da classificação.

Para aplicar métodos de seleção do tipo *Filter*, os atributos são avaliados de acordo com métricas como Chi-square, *Mutual Information*, *Term Stregth*, *Information Gain* (2.3.1), frequência de termos e TF-IDF (2.3.2) e *Gain Ratio* [36][20].

Os principais métodos de avaliação de características utilizados para classificação textual serão apresentados em ordem cronológica.

2.3.1 Information Gain

Conforme apresentado na subseção 2.2.3, *information gain* é uma métrica para descrever quão bem cada característica separa as possíveis classes.

Também conhecida por entropia (H), a quantidade de informação de uma mensagem foi formulada por Claude Shannon em 1948 [37].

A entropia descreve a aleatoriedade na relação entre os atributos e as classes analisadas. O seu cálculo é baseado na quantidade de testes de sim ou não nos valores dos atributos de um exemplo dos dados e ser capaz de predizer a sua classe (como na estrutura de árvore de decisão). Quanto maior a entropia, isto é, quanto mais informação os dados possuírem, mais difícil é predizer a classes, pois são necessários em média mais testes para separação desses dados [13].

Sendo m a quantidade de classes possíveis e p_i a probabilidade de, na partição D [1], ser encontrado um exemplo da categoria C, a quantidade de informação $Info(D)$ necessária para classificar um exemplo de uma classe C é definida pela Equação 2.31 [13].

A quantidade de informação de uma partição pelo atributo A, definida pela Equação 2.32 é quantidade de informação restante para a classificação de um exemplo de D após a separação dos dados pelo atributo. Nessa equação, n é a quantidade de valores discretos possíveis para o atributo $A = \{a_1, a_2, a_3, ..., a_n\}$ e D_j é a quantidade de exemplos em D que apresentam o atributo a_j [13].

A medida de ganho de informação descrita na Equação 2.32 é referente a quantidade de informação obtida por meio da separação dos dados com o atributo A.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2.31}$$

$$Info_A(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \times Info(D) \tag{2.32}$$

$$Gain_A = Info(D) - Info_A(D) \tag{2.33}$$

[1] D é a partição de treinamento dos dados, definido na Seção 2.2.3

Por meio dessa métrica é possível escolher os melhores atributos, isto é, os atributos que descrevem melhor os dados, proporcionando as classes serem preditas posteriormente com menor quantidade de informação [13] [17].

Para dados textuais, o ganho de informação para a classificação dada a presença ou ausência de determinado termo em um documento é calculado para cada termo e são excluídos aqueles que não atingem determinado limite pré-estabelecido [20]

2.3.2 TF-IDF

Na Seção 2.1, a representação de um documento foi definida como o vetor dos termos (2.1) (utilizando *bag-of-words*). É possível também elaborar a representação de documentos com um peso associado a cada termo. Uma técnica para tal representação é o TF-IDF [17].

O artigo de Salton [38] apresenta métodos de representação de termos para sistemas de recuperação de informações. Esses sistemas possuem a aplicação de retornar informações sobre os documentos textuais relacionados através de uma busca de termos, identificação o conteúdo dos documentos do sistema.

Salton descreve as medidas de *recall* e *precision* a fim de determinar a qualidade do sistema de consulta textual, definindo que um bom sistema desse tipo deve produzir alto *recall*, ou seja, retornar todos os resultados relevantes, e também alta *precision*, não retornar resultados inadequados.

Utilizando a representação de um documento em um vetor de termos vezes o seu respectivo peso (Equação 2.34), o estudo do artigo é focado na análise de diferentes ponderações dos pesos dos termos.

$$X = \begin{bmatrix} x_1, w_1 \\ x_2, w_2 \\ x_3, w_3 \\ \vdots \\ x_n, w_n \end{bmatrix} \tag{2.34}$$

Em destaque, as seguintes ponderações são testadas:

- Frequência do termo (TF): É definida pelo número de ocorrências do termo no documento dividido pela quantidade total de termos no documento.
- Frequência inversa do termo (IDF): Sendo N o número total de documentos e N_x o número de documentos nos quais o termo x aparece, o IDF é calculado através da Equação 2.35. Essa formulação indica termos com maior potencial de classificação

dos documentos dentre todos os termos, pois o valor de IDF será menor caso o termo se apresente na maioria dos documentos e maior quando o termo aparecer em poucos documentos [17].

$$IDF = \log_2(\frac{N}{N_x}) \tag{2.35}$$

O autor apresenta três considerações principais acerca da ponderação para ser alcançado alto *recall* e *precision*.

O primeiro fator a ser considerado é a frequência do termo (TF) em um único documento, que aumenta o retorno de resultados relevantes. O segundo fator é a frequência inversa do termo em todos os documentos (IDF), que produz maior precisão ao considerar a ocorrência de um termo no conjunto dos documentos.

Sendo assim, o TF-IDF é definido pela multiplicação da frequência do termo e a frequência inversa do termo (Equação 2.36).

$$\text{TF-IDF} = TF \times IDF \tag{2.36}$$

A terceira consideração é a normalização do tamanho dos vetores dos documentos, aplicada para casos em que os documentos apresentam tamanhos diversos, a fim de não favorecer a comparação com documentos que possuem mais termos. A normalização apresentada por Salton de acordo com a Equação 2.37, em que n é o número total de termos do documento.

$$\frac{t}{\sqrt{\sum_n w_i^2}} \tag{2.37}$$

Para normalizar o vetor completo, cada termo t representado pela quantidade de ocorrências no texto é dividido pelo tamanho do vetor, valor definido pela raiz quadrada da soma dos quadrados dos pesos que representam cada termo [17].

No artigo citado, os experimentos foram conduzidos com cinco coleções de documentos diferentes, o método para ponderação dos termos com os três fatores apresentados produziu o melhor resultado, concluindo que deve ser utilizado como base em comparações com outros métodos de representação e ponderação de termos para sistemas de análise de textos.

É relatado que a representação dos atributos por meio do método TF-IDF permite melhores performances em comparação ao método simplificado que utiliza somente a frequência de termo (TF) [17].

Capítulo 3

Trabalhos Relacionados

Neste capítulo serão elencados trabalhos relacionados a implementações dos modelos de classificação textual e a aplicações de técnicas de extração e seleção de características.

A Figura 3.1 apresenta o grafo de conexões entre artigos produzido pelo aplicativo *connected papers* <https://www.connectedpapers.com/>. Os nós com cor azul claro indicam os artigos citados pela revisão sistemática de Sebastiani [39], incluindo os estudos de Lewis 1994, Yang 1997 e Joachims 1998 apresentados neste capítulo. Os tamanhos dos nós representam a quantidade de citações. Como referência, a revisão citada possui 8104 citações, segundo o sistema do *connected papers*.

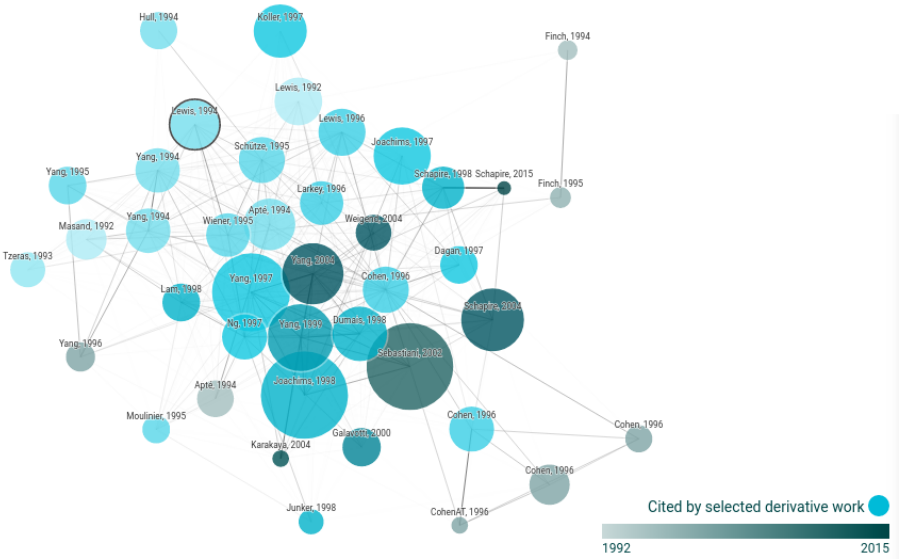

Figura 3.1: Grafo gerado das conexões entre os trabalhos descritos neste capítulo.

Outra conexão entre os trabalhos apresentados é o histórico de citações, começando pelo artigo *Bag of Tricks for Efficient Text Classification* com participação de Tomas Mikolov [21], do centro de pesquisa de inteligência artificial do Facebook. Neste artigo, a implementação de modelo SVM de Joachims de 1998, apresentada ao final deste capítulo, é

citada como um poderoso modelo baseline para a classificação textual. Joachims [18], por sua vez, cita o trabalho de 1997 de Yang e Pedersen [20], ao confirmar a boa performance do modelo kNN na classificação de um mesmo corpus. Por fim, Yang e Pedersen citam o artigo de 1994 de Lewis [40], ao referenciar artigos que utilizam técnicas de redução de dimensionalidade com a extração e seleção de características.

3.1 Avaliação do Modelo de Árvore de Decisão

No trabalho de Lewis [40], *A Comparison of Two Learning Algorithms for Text Categorization*, dois modelos são testados para a tarefa de classificação textual supervisionada em dois corpora em língua inglesa, *Reuters* e MUC3.

O primeiro é um modelo Bayesiano, algoritmo denominado *PropBayes*, que utiliza o teorema de Bayes, explicado na Seção 2.2.1, para determinar a probabilidade de cada exemplo pertencer a uma determinada classe e, então, prediz pela categoria que apresenta maior probabilidade. O segundo modelo é uma árvore de decisão (Seção 2.2.3), algoritmo denominado *DT-min10* [30], que separa os exemplos nas categorias analisadas por meio da métrica de análise dos atributos apresentada na Seção 2.3.1 *Information Gain*.

O corpus Reuters consiste na categorização de 21450 documentos divididos em 14704 para o conjunto de treinamento e 6746 para o conjunto de teste em 135 classes diferentes. Cada documento consiste em um conjunto de 22791 características binárias, representando os termos que ocorrem em dois ou mais documentos do corpus.

O corpus MUC3 é composto por 1500 documentos em classificados em 88 categorias. Cada documento do MUC3 possui 8876 características binárias, representando também os termos que ocorrem em dois ou mais documentos do corpus.

Antes da etapa de classificação, para ambos os algoritmos, o conjunto de dados passou por uma filtragem de características utilizando como métrica para a extração o *Information Gain*. A relevância de cada atributo foi calculada de acordo com a métrica e foi possível testar a performance dos modelos para diferentes quantidades de características, com conjuntos formados pelas n melhores características.

Os resultados de performance do modelo foram avaliados utilizando a métrica *breake-ven* (em português, empate), o maior valor em que as medidas de *precision* e *recall* (Seção 2.1.2) são iguais. O autor relaciona os resultados atingidos em diferentes quantidades de características dos conjuntos de atributos formados utilizando a avaliação do *Information gain*.

O autor aponta que o modelo *PropBayes* sofreu da chamada "maldição da dimensionalidade", ao observar que uma performance máxima foi atingida por um número reduzido de atributos e, ao aumentar a quantidade de atributos, a performance sofre perda. Isso

se dá pelo fato de que muitos atributos adaptam as escolhas do modelo aos dados de treinamento e comete mais erros nos exemplos inéditos do conjunto de teste.

Já a avaliação do modelo *DT-min10* apresenta que a performance continua aumentando até o conjunto com 90 características no corpus *Reuters*, o que demostra a melhor capacidade de avaliação da relevância dos atributos promovida pela técnica de *Informational Gain* no algoritmo de árvore de decisão, condicionada a uma grande quantidade de exemplos de treinamento para os cálculos dos caminhos da árvore serem precisos.

No corpus menor, o MUC-3, a performance possui comportamento semelhante ao *PropBayes*, atingindo um valor máximo para uma quantidade pequena de atributos e perdendo performance com conjuntos com mais atributos.

Essa análise do modelo de árvore de decisão demonstra que esse algoritmo pode ter sua performance aprimorada com o uso de técnicas de filtragem de características (Seção 2.3) anteriormente à aplicação do modelo.

3.2 Avaliação da Relevância dos Atributos de Dados Textuais

Yang e Pedersen [20], no artigo *A Comparative Study on Feature Selection in Text Categorization*, analisa métodos para seleção de características de conjuntos de dados textuais. Esses métodos possuem como finalidade reduzir a dimensionalidade dos textos, selecionando características relevantes para os métodos de classificação ou ainda criando características a partir da combinação das existentes.

O artigo apresenta a alta dimensionalidade do espaço de características como a maior particularidade e dificuldade de representação de dados textuais. Ele aponta que é altamente recomendável reduzir esse espaço, se isso não resultar em diminuição da acurácia, de forma automática, isto é, por meio da aplicação de técnicas de extração de características não manuais, que irão remover os termos não relevantes de acordo com as estatísticas dos termos do corpus e criar novas características que combinam atributos separados.

Os autores usam dois corpora, *Reuters-22173* e *OHSUMED*, para avaliar os seguintes métodos: *Document frequency thresholding* (DF), *Information gain* (IG), *Mutual information* (MI), X^2*-test* (CHI) e *Term strength* (TS).

O método método DF avalia os atributos de acordo com o número de documentos em que um termo aparece. O valor é calculado para todos os termos e são eliminados aqueles que não atingem um limite pré-estabelecido, baseando-se na ideia de que termos raros não possuem informação para predizer as classes em que ocorrem.

O método IG é descrito na Seção 2.3.1. O método MI calcula a relevância de um atributo t, a partir da estimação apresentada na Equação 3.1, em que γ é o número de

vezes que o termo t co-ocorre com a classe C, λ é o número de vezes que o termo t ocorre sem a classe C, v é o número de vezes que a classe C ocorre sem o termo t e N é a quantidade total de documentos.

$$I(t, C) = \log \frac{\gamma \times N}{(\gamma + v) \times (\gamma + \lambda)} \tag{3.1}$$

A métrica CHI calcula a relevância de um atributo t a partir da Equação 3.2. Λ é uma nova variável que representa o número de vezes em que nem o termo t e nem a classe C ocorrem no conjunto de documentos. Segundo os autores, essa métrica mede a falta de independência entre t e C.

$$CHI(t, C) = \frac{N \times (\gamma\Lambda - v\lambda)^2}{(\gamma + v) \times (\lambda + \Lambda) \times (\gamma + \lambda) \times (v + \Lambda)} \tag{3.2}$$

Para as métricas MI e CHI, a relevância de um termo para uma classe calculada por $I(t, C)$ e $CHI(t, C)$, respectivamente, é combinada com a probabilidade de cada classe, condição descrita na Equação 3.3 e a avaliação da classe com a maior relação com a classe é dada pela Equação 3.4, em que $I\_CHI(t, C)$ é $I(t, C)$ ou $CHI(t, C)$, de acordo com o método utilizado.

$$I\_CHI_{avg}(t) = \sum_{i=1}^{m} P_r(C_i) I\_CHI(t, C) \tag{3.3}$$

$$I\_CHI_{max}(t) = \max_{i=1}^{m} I\_CHI(t, C) \tag{3.4}$$

Finalmente, o método TS calcula a relevância de uma atributo t baseando-se na probabilidade de t aparecer em documentos "intimamente relacionados", definidos por pares de documentos que ultrapassam um limite de similaridade calculado por meio do valor similaridade de cosseno.

A métrica de avaliação utilizada pelos autores foi a *precision* média calculada para as predições das classes dos corpora utilizados.

Utilizando os classificadores *k-nearest-neighbor* (KNN) e *Linear Least Squares Fit mapping* (LLSF), os testes conduzidos pelos autores apresentam que que os métodos IG e CHI produzem os resultados mais efetivos, assim como o DF, que possui custo computacional menor, demostrando que os três removeram 90% ou mais termos únicos do texto, em

comparação com os demais métodos testados. Os autores também demonstram uma alta correlação das importâncias dos termos para os três métodos. O método TS foi capaz de reduzir o vocabulário (termos únicos) apenas em 50% e o método MI também apresentou baixa performance, por produzir tendências favorecendo termos raros.

3.3 Avaliação de Modelos de Classificação

Joachims [18], conduz experimentos realizando a classificação em dois corpora, com o objetivo de comparar os resultados de classificação de modelos SVM (Seção 2.2.5) utilizando *kernels* polinomiais e RFB com os seguintes modelos convencionais de classificação: naive Bayes (Seção 2.2.1), k-NN (Seção 2.2.2), o algoritmo Rocchio utilizado em sistemas de recuperação de informação (em inglês, *information retrieval* (IR)) e a árvore de decisão C4.5 (Seção 2.2.3)).

O primeiro corpus, Reuters-21578, possui 9603 documentos no conjunto de treinamento e 3299 no conjunto de teste. Após pré-processamento, o corpus possui 9962 termos distintos no conjunto de treinamento.

O segundo, Ohsumed, possui 10000 documentos para treinamento e 10000 para testes. Após o pré-processamento, o corpus possui 15561 termos distintos no conjunto de treinamento.

Com o objetivo de evitar resultados enviesados pela escolha das características, os modelos citados foram testados com diferentes seleções de atributos. Aplicando o método *Information Gain* (Seção 2.3.1), foram selecionados os 500, 1000, 2000, 5000 e 10000 melhores atributos e os resultados apresentados são correspondentes às seleções com melhor performance.

Avaliando de acordo com a métrica de performance *Precision/Recall* 2.1.2, os modelos SVMs (polinomial e RBF) apresentaram os melhores resultados gerais independente do conjunto de características, demonstrando a capacidade de generalização de SVMs em espaços alta dimensionalidade, eliminando a necessidade da etapa de seleção de características.

Sobre os resultados observados para os métodos convencionais:

- O modelo k-NN apresentou a melhor resultado entre os modelos convencionais;
- Após o k-NN, o algoritmo Rochio apresentou o melhor resultado, seguido do modelo C4.5;
- k-NN, Rochio e C4.5 obtiveram a maior performance utilizando o conjunto de 1000 atributos;

- A melhor performance do modelo Naive Bayes foi utilizando todos os atributos. Ainda, este foi o pior resultado entre as melhores performances dos outros modelos;

Acerca do tempo de execução dos modelos do experimento, apesar das SVMs apresentarem maior custo no tempo de treinamento em comparação aos modelos convencionais *naive Bayes*, Rocchio, k-NN e C4.5, (as SVMs) são mais rápidas do que o modelo k-NN no tempo de classificação.

O artigo de Yang e Pedersen [20] apresenta métodos probabilísticos automáticos que analisam o texto quanto a frequência de seus termos para executar a importante tarefa de redução de dimensionalidade do espaço de atributos dos dados. Utilizando esta lógica de avaliação da importância dos atributos, o descritor implementado neste trabalho é baseado nos valores de frequência dos termos da base textual do Diário Oficial da União.

Em comparação com os estudos apresentados, neste trabalho, será avaliada a performance do algoritmo de Árvore de Decisão para classificação de dados textuais após a utilização de métodos de extração de características implementado para criar um subconjunto de características mais relevantes, método de avaliação de modelos de classificação para dados textuais aplicado por Lewis e Joachins.

Capítulo 4

Metodologia Proposta

Neste capítulo será descrita a metodologia utilizada com o objetivo de classificar, em risco de conluio, publicações do Diário Oficial da União (DOU).

As seções deste capítulo descrevem a base de dados e as etapas de implementação evidenciadas no fluxograma da Figura 4.1: Pré-Processamento, Extração de características, Classificação e Avaliação do modelo de classificação.

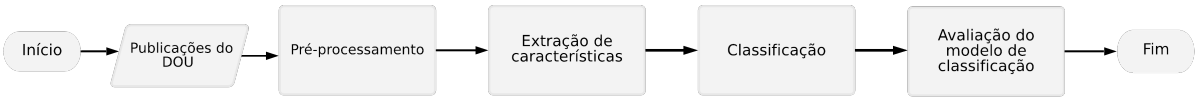

Figura 4.1: Fluxograma da implementação.

As etapas de avaliação do modelo de classificação serão discutidas nas Seções 5.2 e 5.2.2.

4.1 Base de dados: Publicações do DOU

A base de dados utilizada consiste em todas as publicações do Diário Oficial da União de janeiro 1998 até fevereiro de 2020, somando 15.132.968 publicações. A Figura 4.2 apresenta uma página do DOU, referente à Seção 3. O DOU possui ainda mais duas seções, compreendendo [41]:

- Seção 1: "Atos normativos de interesse geral" (leis, decretos, resoluções, instruções normativas, portarias e outros).
- Seção 2: "Atos de pessoal relativos aos servidores públicos"
- Seção 3: "extratos de instrumentos contratuais" (acordos, ajustes, autorizações de compra, contratos, convênios, ordens de execução de serviço, termos aditivos e ins-

EMBRAPA COCAIS

EXTRATO DE TERMO ADITIVO

Espécie: Termo Aditivo; Partes: Embrapa Cocais, CNPJ: 00.348.003/0022-45 e Trivale Administração Ltda, CNPJ: 00.604.122/0001-97; Objeto: Serviços de administração e gerenciamento de fornecimento de combustível; Vigência: 12 meses; Início da Vigência: 02/05/2022. Data Assinatura: 30/04/2021; Vlr. R$ 151.509,85; Signatários: Neusa Alice dos Santos e Aldir Fonseca Lima, pela Embrapa; Vitor Flores de Deus, pela Contratada.

EMBRAPA FLORESTA

AVISO DE LICITAÇÃO
PREGÃO ELETRÔNICO Nº 2/2021 - UASG 135028

Nº Processo: 21175.000426/2021. Objeto: Contratação de empresa(s) para o fornecimento de gases especiais para realização de análises laboratoriais dos projetos de pesquisa da Embrapa Florestas. Total de Itens Licitados: 8. Edital: 11/05/2021 das 08h00 às 17h00. Endereço: Estrada da Ribeira Km 111-cx.p 319 - Patrimônio e Suprimentos-sps, - Colombo/PR ou https://www.gov.br/compras/edital/135028-5-00002-2021. Entrega das Propostas: a partir de 11/05/2021 às 08h00 no site www.gov.br/compras. Abertura das Propostas: 25/05/2021 às 09h00 no site www.gov.br/compras.

REJANE STUMPF SBERZE
Chefe Adjunto de Administração

(SIASGnet - 10/05/2021) 135028-13203-2021NE800027

EMBRAPA HORTALIÇAS

EXTRATO DE COMPROMISSO

Espécie: Termo de Compromisso de Confidencialidade e Outras Avenças, vinculado ao Acordo de Cooperação Técnica e Científica celebrado entre a Embrapa e o CNPq - SAIC/AJU nº 10200.16/0065-2 em 06/07/2016. Partes: Embrapa Hortaliças - CNPJ: 00.348.003/0055-03 e Tadeu Araújo de Souza. Objeto: Permitir ao Bolsista a utilização da infraestrutura da Unidade 212001 na execução de projeto. Valor Global: Sem ônus. Vigência: 17/05/2021 a 30/04/2022. Data de Assinatura: 06/05/2021. Signatários: Warley Marcos Nascimento, Chefe-Geral da Embrapa Hortaliças e Tadeu Araújo de Souza, Bolsista.

EXTRATO DE COMPROMISSO

Espécie: Termo de Compromisso de Confidencialidade e Outras Avenças, vinculado ao Acordo de Cooperação Técnica e Científica celebrado entre a Embrapa e o CNPq - SAIC/AJU nº 10200.16/0065-2 em 06/07/2016. Partes: Embrapa Hortaliças - CNPJ: 00.348.003/0055-03 e Samuel Feitoza Guedes. Objeto: Permitir ao Bolsista a utilização da infraestrutura da Unidade 212001 na execução de projeto. Valor Global: Sem ônus. Vigência: 17/05/2021 a 30/04/2022. Data de Assinatura: 10/05/2021. Signatários: Warley Marcos Nascimento, Chefe-Geral da Embrapa Hortaliças e Samuel Feitoza Guedes, Bolsista.

EXTRATO DE CONTRATO

Espécie: Contrato de validação; Partes: Embrapa Hortaliças, CNPJ nº 00.348.003/0055-03 e Ilso Sebastião Marchesin e outros, CNPJ nº 08.083.498/0001-05. Objeto: integração de esforços entre as partes para implantação de Unidade(S) de Observação - UO, em imóvel de propriedade e/ou posse do COOPERTANTE, no município de São Carlos/SP, visando validar em dois ciclos de cultivo de cinco cultivares de alho comum semi-nobre livres de vírus em sistema orgânico de produção quanto a características de produção e qualidade agronômica e condimentar de bulbos, nas condições e na perspectiva de um produtor comercial de hortaliças orgânicas na região de São Carlos-SP.; Unidade Gestora: Embrapa Hortaliças; Vigência: 07/05/2023; Data da assinatura: 07/05/2021. Signatários: Warley Marcos Nascimento e Henrique Martins Gianvecchio Carvalho, pela Embrapa Hortaliças, e Flavio Roberto Marchesin, pela e Ilso Sebastião Marchesin e outros.

EMBRAPA INSTRUMENTAÇÃO

EXTRATO DE RECONHECIMENTO

Espécie: Instrumento de Reconhecimento de Direitos e Estabelecimento de Obrigações nº nº 23700.21/0014-1; partes: Embrapa Instrumentação e Protech Pesquisa e Desenvolvimento Ltda. Objeto: estabelecer a titularidade de Propriedade Intelectual e Exploração Econômica do(s) Ativo(s) da(s) Tecnologia(s) desenvolvido(s) a partir do desenvolvimento do projeto denominado "Desenvolvimento de máscaras faciais biodegradáveis filtrantes", formal e juridicamente estabelecido no âmbito do "Contrato de Cooperação Técnica" Embrapa Cód. SAIC 23700.21/0013-3. Fonte de recursos: não se aplica; Modalidade de Licitação: não se aplica. Vigência: 10.05.2026; Valor Total: não se aplica Data Assinatura: 10.05.2021. Ass.: José Manoel Marconcini e Débora Marcondes Bastos Pereira (Embrapa); Thiers Massami Uehara (Protech)

EXTRATO DE ACORDO DE COOPERAÇÃO TÉCNICA

Espécie: Acordo de Cooperação Técnica nº 23700.21/0013-3; partes: Embrapa instrumentação e Protech Pesquisa e Desenvolvimento Ltda. Objeto: integração de esforços entre as Partes, para a execução de trabalhos de pesquisa agropecuária, de interesse mútuo, consistente na execução de "Desenvolvimento de máscaras faciais biodegradáveis filtrantes". Fonte de recursos: não se aplica; Modalidade de Licitação: não se aplica. Vigência: 10.11.2022; Valor Total: R$ 157.189,92 (cento e cinquenta e sete mil, cento e oitenta e nove reais e noventa e dois centavos); Data Assinatura: 10.05.2021. Ass.: José Manoel Marconcini e Débora Marcondes Bastos Pereira (Embrapa); Thiers Massami Uehara (Protech)

EMBRAPA MANDIOCA E FRUTICULTURA

AVISO DE LICITAÇÃO
PREGÃO ELETRÔNICO Nº 1/2021 - UASG 135014

Nº Processo: 21186.000486/2021. Objeto: Prestação de serviços de manutenção corretiva nas motobombas e exaustores da Embrapa/CNPMF, com fornecimento de materiais e peças necessários à execução dos serviços, por conta da Contratada. Total de Itens Licitados: 2. Edital: 11/05/2021 das 08h00 às 12h00 e das 14h00 às 17h00. Endereço: Rua Embrapa, S/nº, Centro, Chapadinha - Cruz das Almas/BA ou https://www.gov.br/compras/edital/135014-5-00001-2021. Entrega das Propostas: a partir de 11/05/2021 às 08h00 no site www.gov.br/compras. Abertura das Propostas: 25/05/2021 às 09h00 no site www.gov.br/compras. Informações Gerais: .

PEDRO CANNA BRAZIL RAMOS
Chefe Adj. Administração

(SIASGnet - 10/05/2021) 135014-13203-2021NE000001

EMBRAPA MEIO-NORTE

EXTRATO DE TERMO ADITIVO

Termo Aditivo 04; Partes: Embrapa Meio-Norte, CNPJ n.º 00.348.003/0133-60 e a Empresa Brasil de Comunicação S/A - EBC, inscrita no CNPJ/MF sob o nº 09.168.704/0001-42; Objeto: prorrogação do prazo de Vigência do Contrato Original por mais 12 (doze) meses, com início em 26/05/2021 e término em 26/05/2022; Data da Assinatura: 06/05/2021; Signatários: Flávio Favaro Blanco e Antônio das Graças Lima Filho, pela Embrapa e, Antonio Marinho da Cunha Junior e Ana Carolina Elleres Guedes, pela EBC.

EMBRAPA MILHO E SORGO

EXTRATO DE TERMO ADITIVO

Extrato de Termo Aditivo n° 16. Objeto: Prorrogar a vigência do contrato de locação de imóvel funcional por mais 12 (doze) meses e alterar a cláusula 5ª do contrato sobre o valor do aluguel. Modalidade de Licitação: Não se aplica; Valor Mensal: R$ 76,95. Validade: até 11/05/2022; Data da assinatura: 05/05/2021; Signatários: Roberto Willians Noda - Chefe Adjunto de Administração e Rozemberg Guimarães Arantes - Supervisor do Setor de Patrimônio e Suprimentos da Embrapa Milho e Sorgo, e Paulo Eduardo de Aquino Ribeiro - Locatário.

EXTRATO DE TERMO ADITIVO

Extrato de Termo Aditivo n° 15. Objeto: Prorrogar a vigência do contrato de locação de imóvel funcional por mais 12 (doze) meses e alterar a cláusula 5ª do contrato sobre o valor do aluguel. Modalidade de Licitação: Não se aplica; Valor Mensal: R$ 324,74. Validade: até 14/05/2022; Data da assinatura: 05/05/2021; Signatários: Roberto Willians Noda - Chefe Adjunto de Administração e Rozemberg Guimarães Arantes - Supervisor do Setor de Patrimônio e Suprimentos da Embrapa Milho e Sorgo, e Guilherme Ferreira Viana - Locatário.

EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA

EXTRATO DE ACORDO DE CONFIDENCIALIDADE

Espécie: Acordo de Confidencialidade entre Embrapa e Trion 3D Planning Center. Licitação: Não se aplica. Objeto: Estabelecer a troca e divulgação de informação confidencial entre as partes. Partes: Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA e e Trion 3D Planning Center Vigência: 27/04/2024. Data da assinatura: 27/04/2021. Signatários: Maria Cleria Valadares Inglis e Rafael Vivian, pela Embrapa Recursos. Genéticos e Biotecnologia, Oswaldo Henrique Bastos Salles pela Trion 3D Planning Center

EXTRATO DE ACORDO DE CONFIDENCIALIDADE

Espécie: Acordo de Confidencialidade entre Embrapa e Fertibil Consultoria e Produtos Orgânicos Ltda e Agropecuária Estrela da Manhã Licitação: Não se aplica. Objeto: Estabelecer a troca e divulgação de informação confidencial entre as partes. Partes: Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA e FERTIBIL. CONSULTORIA E PRODUTOS ORGANICOS LTDA e AGROPECUÁRIA ESTRELA DA MANHÃ Vigência: 05/05/2024. Data da assinatura: 05/05/2021. Signatários: Maria Cleria Valadares Inglis e Rafael Vivian, pela Embrapa Recursos Genéticos e Biotecnologia, Maurício Gonçalves Garcia Cid pela Agropecuária Estrela da Manhã e Diego Emerenciano Bringel de Oliveira pela Fertibil Consultoria e Produtos Orgânicos LTDA.

EMBRAPA TABULEIROS COSTEIROS

AVISO DE LICITAÇÃO
PREGÃO ELETRÔNICO Nº 3/2021 - UASG 135013

Nº Processo: 21203000819202101. Objeto: Constitui objeto da presente licitação o registro de preços para aquisição de FORNECIMENTO DE KITs PARA O CAFÉ DA MANHÃ (DESJEJUM), de acordo com as especificações técnicas, condições, quantidades e padrões de desempenho e qualidade estabelecidas no Termo de Referência.. Total de Itens Licitados: 1. Edital: 11/05/2021 das 08h00 às 12h00 e das 13h30 às 17h30. Endereço: Av. Beira Mar, 3250 - Praia 13 de Julho, Jardins - Aracaju/SE ou https://www.gov.br/compras/edital/135013-5-00003-2021. Entrega das Propostas: a partir de 11/05/2021 às 08h00 no site www.gov.br/compras. Abertura das Propostas: 21/05/2021 às 09h00 no site www.gov.br/compras. Informações Gerais: .

PAULO CESAR SILVA DE CARVALHO
Chefe Adjunto de Administração

(SIASGnet - 10/05/2021) 135013-13203-2021NE135013

EMBRAPA UVA E VINHO

EXTRATO DE CESSÃO DE DIREITOS AUTORAIS

Espécie: Termo de Cessão de Direitos Autorais Patrimoniais; Partes: Embrapa Uva e Vinho - CNPJ: 00.348.003/0058-56, Samar Velho da Silveira - CPF nº 627.249.830-68; Cássia Cagliari - CPF nº 036.744.930-71; Elisangeles Baptista de Souza - CPF nº 839.021.129-72; Fátima Miranda D'Ávila Pereira - CPF nº 754.747.350-49; Lucas da Ressurreição Garrido - CPF nº 494.181.146-15; Regis Sivori Silva dos Santos - CPF nº 491.798.530-72 (Cedentes); Objeto: os cedentes cedem à Embrapa, a título gratuito, de forma total e definitiva, em caráter irrevogável e irretratável, nos termos da Lei 9610/1998, de 19.02.1998, os direitos patrimoniais sobre a obra "Grade de Agrotóxicos do Kiwi", doravante designada simplesmente de Obra, decorrentes de sua participação na condição de coautores da obra; Unidade Gestora: 135033; Modalidade de licitação: Não aplicável; Fonte de recurso: Não aplicável; Valor Global: Não aplicável; Data da assinatura: 04.05.2021; Vigência: a partir da data de sua assinatura, pelo prazo do artigo 41 da Lei nº 9.610/98; Signatários: José Fernando da Silva Protas - Chefe-Geral Interino e Adeliano Cargnin, Chefe Adjunto de Pesquisa e Desenvolvimento da Embrapa Uva e Vinho e Samar Velho da Silveira, Cássia Cagliari, Lucas da Ressurreição Garrido, Elisangeles Baptista de Souza, Fátima Miranda D'Ávila Pereira, Regis Sivori Silva dos Santos, cedentes.

AVISO DE LICITAÇÃO
PREGÃO ELETRÔNICO Nº 5/2021 - UASG 135033

Nº Processo: 21206.000553/2021. Objeto: Registro de Preços para aquisição eventual e futura de gases industriais para a Embrapa Uva e Vinho, localizada em Bento Gonçalves-RS.. Total de Itens Licitados: 5. Edital: 11/05/2021 das 08h00 às 11h30 e das 13h00 às 17h30. Endereço: Rua Livramento, 515 Cx. Postal 130 - Bento Gonçalves/rs, Conceição - Bento Gonçalves/RS ou https://www.gov.br/compras/edital/135033-5-00005-2021. Entrega das Propostas: a partir de 11/05/2021 às 08h00 no site www.gov.br/compras. Abertura das Propostas: 21/05/2021 às 09h00 no site www.gov.br/compras. Informações Gerais: .

JOELSIO JOSE LAZZAROTTO
Chefe Adjunto de Administração

(SIASGnet - 10/05/2021) 135033-13203-2021NE000876

Figura 4.2: Página de exemplo do Diário Oficial da União.

trumentos congêneres) editais de citação, intimação, notificação e concursos públicos, comunicados, avisos de licitação entre outros atos da administração pública decorrentes de disposição legal."

Antes de 2018, o DOU era publicado como jornal, em formato PDF. Atualmente, os dados do DOU são publicados no *website* da Impressa Nacional <https://www.in.gov.br/inicio> em formato XML, facilitando a sua captura por meio de API.

Neste trabalho as publicações utilizadas são os atos publicados na Seção 3 do DOU no período mencionado (janeiro 1998 até fevereiro de 2020), referentes às etapas e estados atuais de contratos públicos. Na Figura 4.2, um exemplo de publicação encontrada na base de dados é o tipo "Extrato de Contrato":

> "EXTRATO DE CONTRATO Espécie: Contrato de validação; Partes: Embrapa Hortaliças, CNPJ nº 00.348.003/0055-03 e Ilso Sebastião Marchesin e outros, CNPJ nº 08.083.498/0001-05. Objeto: integração de esforços entre as partes para implantação de Unidade(S) de Observação - UO, em imóvel de propriedade e/ou posse do COOPERTANTE, no município de São Carlos/SP, visando validar em dois ciclos de cultivo de cinco cultivares de alho comum semi-nobre livres de vírus em sistema orgânico de produção quanto a características de produção e qualidade agronômica e condimentar de bulbos, nas condições e na perspectiva de um produtor comercial de hortaliças orgânicas na região de São Carlos-SP.; Unidade Gestora: Embrapa Hortaliças; Vigência: 07/05/2023; Data da assinatura: 07/05/2021. Signatários: Warley Marcos Nascimento e Henrique Martins Gianvecchio Carvalho, pela Embrapa Hortaliças, e Flavio Roberto Marchesin, pela e Ilso Sebastião Marchesin e outros."

Das mais de 15 milhões de publicações, 1907 são rotuladas 'Risco 1', publicações associadas a processos comprovadamente fraudulentos, rotuladas utilizando informações de processos investigativos da Polícia Federal. O restante das publicações é assumido o rótulo 'Risco 0' para a construção do problema, mesmo não sendo possível afirmar que todas essas publicações estão relacionadas a contratos ilibados.

As classes de predição de uma publicação são, portanto, Risco = {'1','0'}, representando se a publicação é similar (1) ou distinta (0) a publicações de contratos com arranjo de conluio.

Devido ao alto desbalanceamento da base em geral, possuindo 98,7% mais exemplos da classe 'Risco 0', foram criados 100 conjuntos balanceados, isto é, com 1907 publicações Risco 1 e 1907 Risco 0, estas selecionadas aleatoriamente da base de dados para cada um dos 100 conjuntos, excluindo aquelas já rotuladas 'Risco 1'. Estes 100 conjuntos formam a base de dados final, utilizada para classificação (Seção 4.4.1).

4.2 Pré-Processamento

No pré-processamento foram retirados os caracteres referentes a marcações HTML e as *stop-words*. Foi utilizada a lista de *stop-words* da biblioteca Spacy para a língua portuguesa.

A próxima etapa do pré-processamento foi a lemmatização, processo semelhante ao *stemming* descrito na Seção 2.1.1, com a diferença de que, na técnica de *stemming*, a transformação dos termos é resultante da simples remoção do sufixo dos termos sem considerar o contexto de utilização, enquanto a técnica de lemmatização retira a flexões dos termos produzindo termos raízes que realmente existem na língua em questão [42].

Posteriormente, os termos foram rotulados de acordo com a sua semântica no texto, utilizando as chamadas *Universal POS (part-of-speech) tags*, que são rotulações designadas aos termos de acordo com sua função gramatical no texto. Foram selecionados somente os termos rotulados nomes próprios (PROPN), substantivos (NOUN), advérbios (ADV), adjetivos (ADV) e verbos (VERB).

Para finalizar a etapa de pré-processamos dos dados, os símbolos e acentuações dos texto foram removidos.

4.3 Extração de Características

Foi utilizada a representação *bag-of-words*, isto é, cada publicação, equivalente à instância 'documento' descrita no Capítulo Fundamentação Teórica, foi considerada uma lista de palavras únicas da publicação com a contagem da ocorrências dos termos, que são as ditas características ou atributos de uma publicação.

A contagem simples foi substituída pela frequência do termo na publicação, desta forma, sendo A o conjunto de atributos $\{t_1, t_2, t_3, ..., t_n\}$, a frequência do termo t_n é a sua quantidade de ocorrência na publicação dividido pela quantidade total de termos na publicação.

As publicações apresentam diferentes tamanhos, variando de dezenas a milhares de termos. Com o objetivo de não influenciar a comparação com o tamanho das publicações, foi aplicada uma normalização max-min (Equação 2.16) aos valores de frequência. Nessa normalização, os valores associados a cada termo são mapeados para o intervalo $[1, 0]$, sendo 1 o valor associado ao(s) termo(s) mais frequente(s) e 0 o valor associado ao(s) termo(s) menos frequente(s) na publicação em questão. Os valores intermediários são calculados por meio da proporção.

Para avaliação visual da associação dos valores de frequência aos termos, foi criada uma estrutura ordenada de níveis de frequência. A estruturação consiste em separar os

termos de acordo com níveis de frequência em intervalos de 5%, a partir dos valores normalizados. Como exemplo, será utilizado texto após o pré-processamento de uma publicação:

"prefeitura municipal rios extratos contratos contrato tectenge tecnologia inscrito ojeto contratacao grupos formais informais agricultura familiar empreendedores familiares rurais visando aquisicao generos alimenticios preparacao alimentacao escolar adquiridos diretamente agricultura familiar objetivando atendimeto lei resolucoes desenvolvimento programa nacional alimentacao ecolar unidades escolares rede municipal ensino mnicipio rios ba vigencia contrato ailton correia martins inscrito objeto cotratacao empresa fornecimento mobiliario atender necessidades municipio rios bahia vigencia contrato b s silva epp inscrito cnpj objeto contratacao empresa especializada imunizacao controle pragas urbanas atvidade quimica dedetizacao descupinizacao desinsetizacao lipeza desinfeccao quimica desincrustacao reservatorio caixas agua predios anexo ii unidades municipio rios bahia vigencia"

Seguindo o exemplo, a estrutura de níveis criada é descrita na Tabela 4.1.

Tabela 4.1: Representação da estrutura de níveis ordenados

Nível de frequência	**Tupla do termo e sua respectiva frequência normalizada**
[1, >0,95]	[('rios', 1.0)]
[0,95, >0,90]	[ ]
[0,90, >0,85]	[ ]
[0,85, >0,80]	[ ]
[0,80, >0,75]	[ ]
[0,75, >0,70]	[ ]
[0,70, >0,65]	[('contrato', 0.6666666666666666), ('inscrito', 0.6666666666666666), ('vigencia', 0.6666666666666666)]
[0,65, >0,60]	[ ]
[0,60, >0,55]	[ ]
[0,55, >0,50]	[ ]
[0,50, >0,45]	[ ]
[0,45, >0,40]	[ ]
[0,40, >0,35]	[ ]
[0,35, >0,30]	[('municipal', 0.3333333333333337), ('contratacao', 0.3333333333333337), ('agricultura', 0.3333333333333337), ('familiar', 0.3333333333333337), ('alimentacao', 0.3333333333333337), ('unidades', 0.3333333333333337), ('objeto', 0.3333333333333337), ('empresa', 0.3333333333333337), ('municipio', 0.3333333333333337), ('bahia', 0.3333333333333337), ('quimica', 0.3333333333333337)]
[0,30, >0,25]	[ ]
[0,25, >0,20]	[ ]
[0,20, >0,15]	[ ]

Os 3 níveis com menor frequência foram descartados, com o objetivo de redução das características dos dados. O limiar foi determinado por meio da observação do gráfico de dispersão dos valores de frequência normalizados em análise aos seus respectivos rótulos (Figura 4.3).

Para a criação deste gráfico, foi utilizado um subconjunto balanceado dos dados, isto é, 1907 publicações Risco 1 e 1907 Risco 0. As publicações Risco 0 foram escolhidas aleatoriamente dentre os exemplos de Risco 0 do banco de dados completo.

Neste gráfico, portanto, os pontos vermelhos e verdes são todos os termos do subconjunto balanceado, indicando se o termo pertence a uma publicação Risco 1 ou Risco 0, respectivamente. A sobreposição de muitos termos com valores pertencentes aos 3 níveis de menor frequência indicou a baixa relevância desses termos para a categorização.

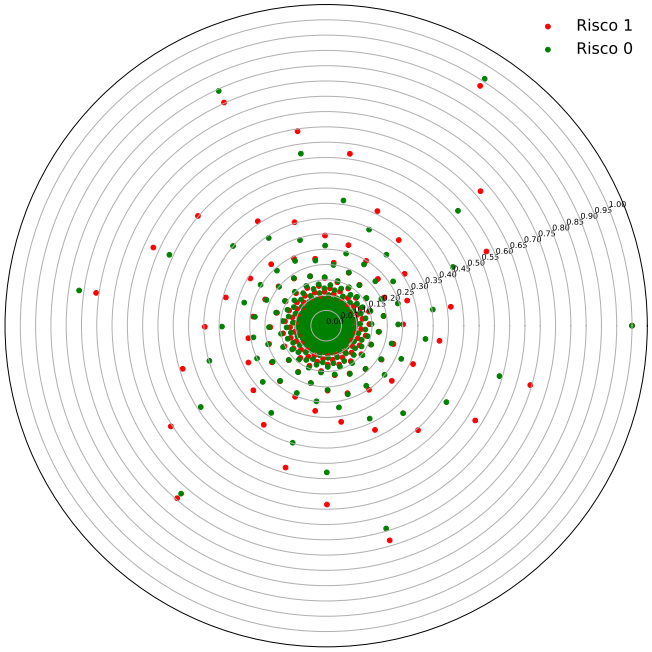

Figura 4.3: Gráfico de dispersão dos termos. Os valores dos anéis do gráfico são os limiares dos níveis de frequência em intervalos de 5%. Os pontos vermelhos representam os termos do vocabulário que pertencem a publicações rotuladas 'Risco 1'. Os pontos verdes representam os termos do vocabulário que pertencem a publicações rotuladas 'Risco 0'.

O gráfico após a retirada dos 3 menores níveis (Tabela 4.1) do subconjunto descrito é observado na Figura 4.4. Neste gráfico também foram apresentados os termos correspondentes a cada ponto.

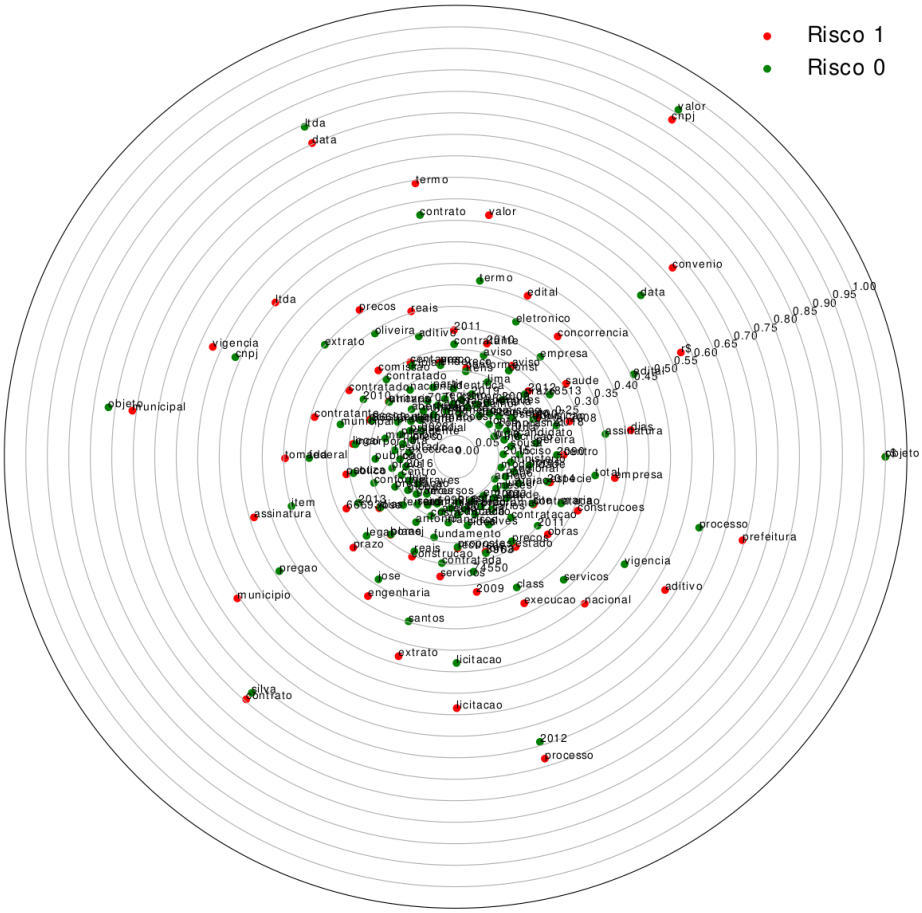

Figura 4.4: Gráfico de dispersão dos termos com a retirada dos 3 níveis de menor frequência e apresentação dos termos associados aos pontos.

Essa estrutura ordenada de níveis com remoção dos níveis de menor frequência foi a estrutura utilizada como descritor das características de uma publicação.

Para ser utilizada com entrada do algoritmo de classificação, cada publicação teve seus termos discriminados de acordo com suas frequências normalizadas na estrutura descrita. Em uma estrutura de dataframe, cada publicação p_N passou a ser representada pela combinação dos valores de frequência f atribuídos a cada termo t_n. Nota-se que muitos campos da estrutura não serão preenchidos e estes foram determinados NaN.

4.4 Classificação

4.4.1 Conjuntos de validação cruzada

O conjunto de validação cruzada utilizado contém 100 conjuntos balanceados, contendo as 1907 publicações Risco 1 e 1907 Risco 0, conforme descrito na Seção 4.1.

	t_1	t_2	t_3	...	t_n
p_1	f	f	f	...	f
p_2	f	f	f	...	f
p_3	f	f	f	...	f
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
p_N	f	f	f	...	f

Tabela 4.2: Estrutura de descrição das publicações por meio das frequências dos seus termos

Cada conjunto consiste em um arquivo em formato JSON, que descreve uma estrutura em dicionário. A primeira chave do dicionário é referente a separação das publicações em conjuntos de treino, 80%, e teste, 20%. Cada publicação desses conjuntos possui as chaves descritas na Tabela 4.3.

Tabela 4.3: Estrutura de uma publicação de treinamento ou teste da base de dados.

Chave	**Valor**
risco	0 ou 1
txt	Texto puro da publicação (antes de qualquer pré-processamento)
id	Chave de identificação da publicação
date	Data de publicação no DOU

4.4.2 Treinamento do modelo de classificação

Foi utilizado o modelo de classificação *Decision Tree*, explicado na seção 2.2.3, com o índice gini como função de avaliação de qualidade da separação dos dados.

Para a entrada deste algoritmos, as publicações do conjunto de treinamento de cada conjunto de validação cruzada foram estruturadas conforme apresentado na Tabela 4.2, acrescentando os seus respectivos rótulos, apresentados na coluna "Risco"da Tabela 4.4.

	t_1	t_2	t_3	...	t_n	Risco
p_1	f	f	f	...	f	0 | 1
p_2	f	f	f	...	f	0 | 1
p_3	f	f	f	...	f	0 | 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
p_N	f	f	f	...	f	0 | 1

Tabela 4.4: Estrutura de descrição das publicações com o respectivo rótulo na última coluna.

Para cada um dos 100 conjuntos de validação cruzada, a estrutura de *dataframe* foi dividida no conjunto X, que possui as linhas das publicações descritas por suas características, as frequências dos termos das colunas, e no vetor y, que é a coluna 'Risco'. Essa

divisão da estrutura foi feita para o conjunto de treinamento e de teste, em proporção de 80% e 20%, produzindo assim os conjuntos: X de treino, X de teste, y de treino e y de teste.

O algoritmo de árvore de decisão recebeu como entrada os conjuntos de X e y de treinamento para calcular, por meio do índice gini, a separação dos dados e, consequentemente, os caminhos da árvore de decisão que resultam nas classes determinadas. Seu retorno, portanto, foi o estimador de predições ajustado por meio da árvore de decisão resultante.

Os conjuntos X e y de teste foram utilizados exclusivamente para os testes de performance em cada conjunto de validação cruzada, com o objetivo de basear a avaliação do modelo de classificação nos resultados de predições para exemplos inéditos. Os resultados dos testes são discutidos nas Seções 5.2.1 e 5.2.2.

Devido ao objetivo deste trabalho ser o teste da estrutura de descritor implementada, os parâmetros foram mantidos com os valores padrões descritos na documentação da biblioteca utilizada *scikitlearn* [43] para o algoritmo de árvore de decisão, a fim de possibilitar a comparação de performance com resultados obtidos de classificadores desenvolvidos como *baseline* do projeto.

Capítulo 5

Resultados

Neste capítulo serão discutidos os resultados obtidos na implementação proposta.

5.1 Tecnologias Empregadas no Desenvolvimento da Implementação

A implementação deste trabalho foi feita utilizando a linguagem de programa python versão 3.7.7, principalmente por meio da interface de programação *Jupyter Notebook* [44]. As bibliotecas utilizadas no contexto geral da implementação estão listadas na Tabela 5.1.

Os testes conduzidos neste trabalhos foram executados por meio de acesso remoto no servidor montado na Universidade de Brasília para o projeto *Deep Vacuity*, armazenando os resultados obtidos neste servidor com cópias locais. As especificações do computador são:

- Processador AMD Ryzen Threadripper 3970X Cache 128MB
- 32 CPUs (64 vCPUs)
- 64 GB RAM (2666 MHz)
- 980 GB Armazenamento SSD
- 4TB Armazenamento HDD (4 discos de 1TB)
- 2 interfaces de rede 1Gbit/s
- 16GB GPU Gforce 2080 (2 placas)

Tabela 5.1: Lista de tecnologias empregadas na implementação.

Módulo Empregado	Função	Nome	Descrição	Justificativa	Versão
Biblioteca json	Interação com arquivos em formato .json	json	Biblioteca que permite ações de ler e escrever arquivos .json	Os arquivos da base de dados deste trabalho estão em formato .json.	2.0.9
Biblioteca pandas	Criação e interação de estruturas em formato *Dataframe*	pandas	Biblioteca para estruturação e análise de dados	Utilizada para organização dos dados antes e após o pré-processamento e para a criação das estruturas no formato correto de entrada para o método de classificação.	1.2.1
Biblioteca Numpy	Criação e interação de estruturas em formato de vetores	numpy	Biblioteca para estruturação e análise de dados	Utilizada para manipulação das estruturas representativas dos documentos.	1.18.5
Biblioteca Spacy	Método de definição lista de *stop-words*	spacy	Biblioteca para processamento de linguagem natural	Utilização para definição das *stop-words* para a língua portuguesa.	2.3.4
Módulo collections	Método para a implementação da representação *bag-of-words*	collections	Módulo para determinação da quantidade de termos únicos no texto	Utilização para contagem dos termos únicos dos documentos.	(python) 3.7.7
Módulo tree	Módulo da biblioteca *scikit learn* para aplicação do algoritmo de árvore de decisão para classificação	tree	Módulo para a tarefa de classificação por meio de árvores de decisão	Utilização na implementação da etapa de classificação	0.23.2
Módulo metrics	Módulo da biblioteca *scikit learn* para avaliação dos resultados	metrics	Módulo com métodos para o cálculo de diferentes métricas	Utilizado para cálculo das métricas de avaliação de performance acurácia, *f1-score*, matriz de confusão e curva ROC.	0.23.2
Módulo matplotlib.pyplot	Interface para a biblioteca matplotlib	matplotlib .pyplot	Módulo para a criação de gráficos e figuras em python	Utilizado para a criação dos gráficos de dispersão e para a apresentação das curvas ROC.	(python) 3.7.7
Módulo os	Interface para interação com o sistema operacional	os	Métodos para ler, escrever, abrir e manipular arquivos e interação com diretórios	Utilizado para interagir com os diretórios de arquivos do projeto	(python) 3.7.7
Módulo SimpleImputer	Módulo da biblioteca *scikit learn* para transformação de dados faltantes (NaN)	sklearn.impute.SimpleImputer	Implementa o método para lidar com campos dos dados com valores NaN	Utilizado para tratar a estrutura *dataframe* na qual os dados foram organizados (Tabela 4.2)	0.23.2

5.2 Avaliação dos Resultados

Nesta seção, serão apresentados os resultados obtidos pela metodologia proposta no Capítulo 4. Devido a solução proposta ser um processamento sequencial e incremental, optou-se em apresentar cada conjunto entrada e saída de cada etapa de forma individual, a partir de avaliação por ablação. Sendo que também foi optado em apresentar os resultados organizados em quantitativos e qualitativos da metodologia implementada, a ser apresentado nas seções a seguir.

5.2.1 Avaliação Quantitativa

Avaliação quantitativa da etapa de pré-processamento

Foram calculados os tamanhos dos textos de todas as publicações dos 100 conjuntos de validação cruzada antes e após o pré-processamento.

A quantidade de atributos das publicações varia de dezenas a até milhares, porque no DOU também são publicados editais completos. A menor publicação possui 64 palavras e a maior possui 30668, com a média de aproximadamente 913 palavras. Após o pré-processamento, as publicações apresentaram, em média, 452 palavras, diminuição da média de aproximadamente 61% da quantidade de palavras de uma publicação.

Avaliação quantitativa da etapa de extração de caraterísticas

A estrutura ordenada de níveis apresentada na Seção 4.3 como método criado para extração de característica apresentou redução expressiva do conjunto de atributos dos dados.

Como comparação, em um conjunto balanceado tomado como exemplo, antes da metodologia aplicada de extração de características, cada publicação era representada por 17755 termos únicos após o pré-processamento. A estrutura implementada descreve as mesmas publicações do conjunto de exemplo com 3273 atributos, apresentando redução em 81% do espaço de características, que são os termos únicos.

Avaliação quantitativa da etapa de classificação

O resultado da classificação foi calculado por meio das métricas acurácia e *f1-score macro*, métricas descritas na Seção 2.1.2.

Os resultados obtidos de média das métricas para os 100 conjuntos balanceados de validação são apresentados na Tabela 5.2.

No momento de implementação da etapa de extração de características 2.3, foi utilizado um conjunto balanceado diferente dos 100 conjuntos de validação utilizados no

	Média	Variância
Acurácia	74,98%	1,70%
F1 score	74,94%	1,71%

Tabela 5.2: Média dos 100 conjuntos balanceados e desvio padrão das métricas acurácia e *f1-score*.

experimento final. Esse conjunto em particular apresentou melhores resultados, por isso esperavam-se melhores resultados gerais.

Entretanto, é comum que o conjunto no qual o tratamento de dados foi elaborado apresente melhores resultados, pois os ajustes foram feitos para esse conjunto em específico. É recomendado que os ajustes de pré-processamento e treinamento não sejam feitos considerando em primeiro momento todo o conjunto de dados disponíveis para prevenir a super adaptação do modelo aos dados, evitando resultados muito diferentes para exemplos inéditos.

Ainda assim, os resultados apresentados foram bons, considerando que a etapa de extração e seleção de característica reduziu em 81% a dimensionalidade em um conjunto de exemplo de validação.

A matriz de confusão de um conjunto balanceado ('dic_raw_6_1') é apresentada na Tabela 5.3. Esse conjunto foi escolhido para apresentação da matriz de confusão de todos os conjuntos por apresentar valores de acurácia e f1-score, 75,13% e 75,12%, respectivamente, próximos aos valores de média.

A matriz de confusão apresenta a proporcionalidade entre as quantidades de predições corretas, verdadeiros positivos (VP) e verdadeiros negativos (VN), e predições erradas, falsos positivos (FP) e falsos negativos (FN).

Para esse conjunto, é demonstrado por meio da matriz de confusão que a predição obtida com o modelo de classificação implementado apresenta proporções de acerto e erro similares para as duas classes. Isso quer dizer que as duas classes são classificadas com taxas de erro similares. Uma avaliação posterior é a avaliação de importância para os dois tipos de erros. Neste caso, os erros foram mantidos balanceados.

Classes	Resultado das predições	
Risco 0	38,35% (VP)	13,21% (FN)
Risco 1	11,64% (FP)	36,78% (VN)

Tabela 5.3: Matriz de confusão do conjunto 'dic_raw_6_1'.

5.2.2 Avaliação Qualitativa

Avaliação qualitativa da etapa de pré-processamento

O pré-processamento é demonstrado no exemplo abaixo:

Publicação:

> "WELLINGTON ANTONIO RODRIGUES DE OLIVEIRA Gestor PREFEITURA MUNICIPAL DE JOÃO PESSOA EXTRATO DE TERMO ADITIVO CONCORRÊNCIA PÚBLICA $N°$ 06/2013/SEPLAN INSTRUMEO: 7° Termo Aditivo ao Contrato $n°$ 01/2014/SEPLAN, assinado em 01/04/2016. PARTES: Prefeitura Municipal de João Pessoa e a COECC Engenharia, Comércio e Construções Ltda., OBJETO: Cotratação de Empresa Especializada para Execução dos Serviços de Reabilitação da Lagoa do Parque Solon de Lucena da Cidade de João Pessoa. FINALIDADE: É objeto é a prorrogação de prazo por mais 90 (noventa) dias corridos. SIGNATÁRIOS: Cássio Augusto Cananéa Andrade /PMJP e a Sr. Eduardo Ribeiro Victor/COMPECC."

Publicação após o pré processamento:

> "wellington antonio rodrigues oliveira gestor prefeitura municipal termo aditivo concorrencia publica n seplan instrumeo termo aditivo contrato n assinado partes prefeitura municipal joao pessoa objeto cotratacao empresa especializada execucao servicos reabilitacao lagoa parque solon lucena cidade joao pessoa finalidade objeto prorrogacao prazo corridos signatarios"

Notou-se que os textos originais das publicações são ruidosos, ao apresentarem erros de grafia nas palavras devido as publicações antigas serem provenientes de digitalização em baixa qualidade dos documentos físicos do Diário Oficial da União.

Foram retirados do texto original pontuações, símbolos e acentuações, além da normalização em letras minúsculas. Também foram retiradas as *stop-words*, como demonstrado no exemplo, o texto após o pré-processamento não possui termos como 'é', 'a', 'de', 'do' e 'ao'. O nome pessoal 'wellington antonio rodrigues oliveira' foi mantido, conforme apresentado na metodologia proposta para a etapa de pré-processamento (Seção 4.2).

Avaliação qualitativa da etapa de extração de caraterísticas

O gráficos de dispersão dos termos apresentados na Seção 4.3 para desenvolvimento da etapa de extração de características indicou a divisão visual dos termos das classes 'Risco 1' e 'Risco 0'.

Com a observação do gráfico, também foi possível avaliar a relevância dos termos de acordo com os níveis de frequência. Os termos dos três níveis de menor frequência puderam ser removidos do vocabulário e da representação de um documento.

Houve ainda leve aumento de performance da classificação considerando a métrica de *f1-score* com a diminuição da dimensionalidade dos atributos em comparação com os resultados do mesmo algoritmo de Árvore de decisão no trabalho de classificadores *baseline* desenvolvidos para esse projeto. Essa implementação utilizou os mesmos conjuntos de dados, com condições idênticas de pré-processamento, porém sem a etapa de extração de características, e apresentou *f1-score* médio de 73,76%.

Avaliação qualitativa da etapa de classificação

A curva ROC e o valor AUC apresentados na Figura 5.1 são resultantes da aplicação do modelo de classificação proposto no conjunto 'dic_raw_6_1', mesmo conjunto da matriz de confusão apresentada na Seção 5.2.1 Avaliação quantitativa da etapa de classificação.

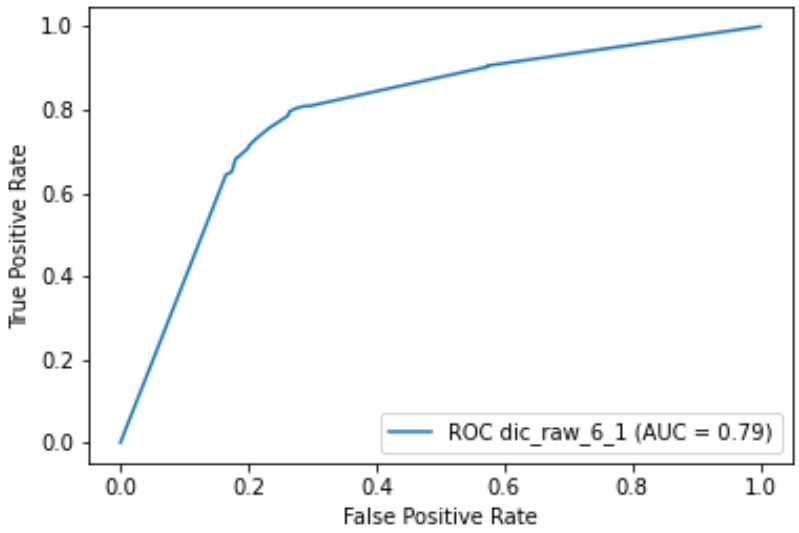

Figura 5.1: Curva ROC e valor AUC calculados para as predições resultantes da classificação no conjunto balanceado 'dic_raw_6_1'.

Conforme considerações da Seção 2.1.2 Curva Característica de Operação do Receptor (ROC) e AUC, a curva apresenta o formato desejável, localizando-se na parte superior esquerda do espaço definido pelos eixos de taxa de falsos positivos (TFP), *False Positive Rate*, eixo x, e taxa de verdadeiros positivos (TVP), *True Positive Rate*, eixo y.

A curva apresenta valores relativamente altos para a taxa de verdadeiros positivos e também relativamente baixos para a taxa de falsos positivos, indicando que as predições do modelo foram consideravelmente boas.

O valor da área abaixo da curva (AUC) confirma a boa qualidade da classificação dos exemplos de teste ao aproximar-se do valor ideal 1. O menor valor de AUC atingindo foi 0,74 e o maior valor foi 0,82, curvas apresentadas na Figura 5.2.

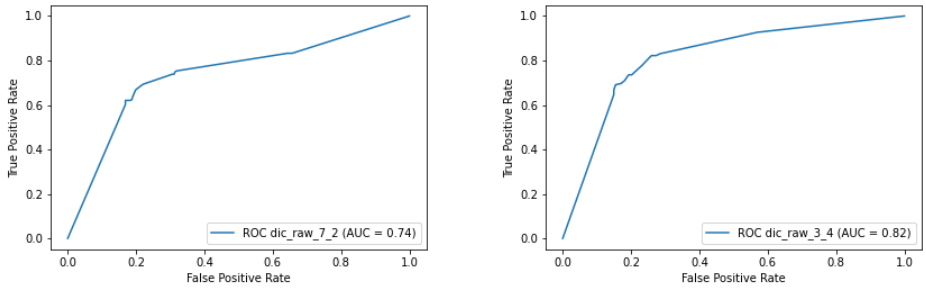

Figura 5.2: Curva ROC e valor AUC resultantes das predições dos conjuntos balanceados 'dic_raw_7_2' (esquerda) e 'dic_raw_3_4' (direta).

5.2.3 Discussão Geral dos Resultados

Uma outra série de procedimentos mais simples para a etapa de pré-processamento foi implementada, sem a ocorrência da *lemmatização* e seleção dos termos segundo sua função gramatical no texto. Compreendia somente a retirada de *stop-words* da língua portuguesa, de sufixos referentes a ênclises de pronomes e de símbolos como barras, pontuações e acentuações, além da filtragem para retirada de todas os termos com menos de três letras, exceto siglas dos estados brasileiros.

Este pré-processamento simples foi utilizado no conjunto balanceado formado para a elaboração estrutura de extração de características (alheio aos 100 conjuntos de validação cruzada), que apresentou melhores resultados de *f1-score* médio e acurácia para as predições, conforme apresentado na seção 5.2.1. Entretanto, foram utilizados os procedimentos descritos na seção 4.2 a fim de manter a padronização dos processos seguidos para a obtenção dos resultados dos classificadores *baseline*, descrita na Seção 5.2.2.

O pré-processamento diverso inclui-se nas possibilidades dos motivos pelos quais o teste em larga escala apresentou resultados diferentes aos resultados do conjunto utilizado na etapa de extração de características. Todavia, conforme discutido na Seção 5.2.1, é provável que esse comportamento tenha sido observado devido ao comportamento natural de obtenção de resultados mais baixos para predições de exemplos diversos aos quais os ajustes das etapas em pré-classificação foram feitos, isso porque a compreensão dos dados foi realizada em um conjunto de dados em específico. Ou seja, as observações dos dados nos gráficos de dispersão apresentados, que levaram às seleções dos níveis de frequência de corte e todas as análises sobre os dados foram influenciadas por esse conjunto de dados.

A comparação de resultados do mesmo algoritmo de classificação, a árvore de decisão, utilizando os mesmos dados resultantes do pré-processamento igual, com a única diferença sendo a aplicação da estrutura de descritor implementada neste trabalho, possibilita a afirmação que o método de extração de característica proposto foi eficiente para

aprimorar a performance da classificação, além de reduzir consideravelmente a quantidade de atributos necessários para descrever as publicações.

Também para manter os procedimentos padrões para a comparação, nenhum dos parâmetros possíveis de ajuste do algoritmo de árvore de decisão implementado pela biblioteca *scikitlearn* [43] foi alterado. Estes parâmetros podem ser mais explorados para aperfeiçoar os resultados obtidos. Outro ajuste possível são combinações diferentes de níveis de frequência na estrutura do descritor, sendo possível encontrar seleções que caracterizem melhor os dados.

Capítulo 6

Conclusões e Trabalhos futuros

O estudo de diferentes técnicas de classificação, extração e seleção de características, bem como a análise dos desenvolvimentos aplicados pelos autores citados nos capítulos de Fundamentação Teórica e Trabalhos Relacionados forneceram o entendimento dos processos aplicados para a tarefa de classificação de dados textuais.

Avaliando os objetivos específicos estabelecidos no Capítulo 1, uma nova abordagem para o desenvolvimento de descritores para bases textuais foi proposta, implementada e testada. A estrutura de níveis de frequência devolvida para avaliar a relevância dos termos e reduzir a dimensionalidade do espaço de atributos foi testada com o modelo de classificação de árvore de decisão.

Os resultados quantitativos e qualitativos apresentados para a etapas de pré- processamento e extração de características foram satisfatórios na medida em que proporcionaram a redução de dimensionalidade esperada e desejada em bases textuais, pois, como apontado por diversos autores citados, a alta quantidade de características é o maior desafio do trabalho de classificação de dados textuais.

Apesar do teste em todos conjuntos de validação apresentar resultados menores de performance do que os atingidos com o conjunto utilizado para a elaboração e ajustes da estrutura de extração de características, a avaliação dos resultados da classificação das publicações em risco de conluio utilizando o método de extração de características desenvolvido apresentou resultados consideravelmente bons, conclusão possível a partir da análise dos valores de acurácia e *f1-score*, além das curvas ROC obtidas com os resultados das predições.

Destaca-se ainda que o processo de classificação realizado com a utilização da técnica de extração de caraterísticas apresentou *f1-score* médio maior do que o mesmo processo de classificação realizado sem a utilização da técnica, resultado descrito na seção 5.2.2.

Diante do exposto, conclui-se que o objetivo geral deste trabalho foi cumprido, ao apresentar um método eficaz de extração de características, utilizado junto a um modelo

de classificação para a avaliação de risco de conluio das publicações do Diário Oficial da União.

6.1 Trabalhos Futuros

Para a análise do descritor implementado, trabalhos futuros são pautados para a sua utilização com modelos de redes neurais desenvolvidos para este projeto que, em comparação ao modelo mais simples de árvore de decisão utilizado neste trabalho, já apresentam melhores resultados, apontados no artigo *"Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach"* [12]. É esperado que o descritor implementado colabore para o aprimoramento dos resultados.

Trabalhos futuros também incluem o aprimoramento do descritor implementado, testando diferentes seleções de níveis de frequência da estrutura ordenada e a influência dessas seleções nos resultados finais de classificação.

Referências

[1] Cobran, Daniel e David Banys: *AUC (Area under the ROC Curve)*, 2020 (acessado em Maio 8, 2021). https://docs.paperspace.com/machine-learning/wiki/auc-area-under-the-roc-curve. ix, 11

[2] *Distribuição Normal (Gaussiana)*, (acessado em Abril 25, 2021). https://www.inf.ufsc.br/~andre.zibetti/probabilidade/normal.html. ix, 13

[3] Navlani, Avinash: *KNN Classification using Scikit-learn*, (acessado em Abril 25, 2021). https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. ix, 14

[4] Navlani, Avinash: *Decision Tree Classification in Python*, (acessado em Abril 25, 2021). https://www.datacamp.com/community/tutorials/decision-tree-classification-python. ix, 15

[5] Weisstein, Eric W.: *Slope-Intercept Form*, (acessado em Abril 25, 2021). https://mathworld.wolfram.com/Slope-InterceptForm.html. ix, 19

[6] *Logistic function*, (acessado em Abril 25, 2021). https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic.html#sphx-glr-auto-examples-linear-model-plot-logistic-py. ix, 20

[7] Navlani, Avinash: *Support Vector Machines with Scikit-learn*, (acessado em Abril 25, 2021). https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python. ix, 21

[8] documentation scikit-learn online: *SVM-Kernels*, (acessado em Maio 8, 2021). https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html. ix, 22

[9] Cayan Portela; Fernanda Amorim, Gustavo Monteiro; Jader Martins e Mariana Montenegro: *Tutorial de SVM*, (acessado em Maio 8, 2021). https://lamfo-unb.github.io/2017/07/13/svm/. ix, 22

[10] Conhecimento Livre, Open Knowledge Brasil Rede pelo: *serenata-de-amor*. https://github.com/okfn-brasil/serenata-de-amor, 2018. 2, 4

[11] Oliveira Lopes, Alan de: *O Efeito Pedagógico de Operações da Polícia Federal: Um Estudo de Caso da Operação"Caixa de Pandora"*. Revista Brasileira de Ciências Policiais, 6(1):67–85, 2015. 2, 4

[12] Lima, Marcos, Roberta Silva, Felipe Lopes de Souza Mendes, Leonardo R. de Carvalho, Aleteia Araujo e Flavio de Barros Vidal: *Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach*. Em *Findings of the Association for Computational Linguistics: EMNLP 2020*, páginas 1580–1588, Online, novembro 2020. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.findings-emnlp.143. 2, 4, 52

[13] Han, Jiawei, Jian Pei e Micheline Kamber: *Data mining: concepts and techniques*. Elsevier, 2011. 2, 5, 6, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 24, 25

[14] Santos Nakamura, André Luiz dos: *A infraestrutura e a corrupção no Brasil*. Revista Brasileira de Estudos Políticos, 2018. 3

[15] Brasil: *Lei n. 8.666, de 21 de junho de 1993*, 1993 (Acessado Dezembro 16, 2020). http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm. 4

[16] Brasil: *Lei n. 10.520, de 17 de julho de 2002*, 2002 (Acessado Dezembro 16, 2020). http://www.planalto.gov.br/ccivil_03/LEIS/2002/L10520.htm. 4

[17] Bramer, Max: *Principles of data mining*, volume 180. Springer, 2007. 5, 6, 7, 8, 9, 14, 25, 26, 27

[18] Joachims, Thorsten: *Text categorization with support vector machines: Learning with many relevant features*. Em Nédellec, Claire e Céline Rouveirol (editores): *Machine Learning: ECML-98*, páginas 137–142, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg, ISBN 978-3-540-69781-7. 5, 22, 29, 32

[19] Wang, Sida I e Christopher D Manning: *Baselines and bigrams: Simple, good sentiment and topic classification*. Em *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, páginas 90–94, 2012. 6

[20] Yang, Yiming e Jan O Pedersen: *A comparative study on feature selection in text categorization*. Em *Icml*, volume 97, páginas 412–420. Nashville, TN, USA, 1997. 7, 24, 25, 29, 30, 33

[21] Joulin, Armand, Edouard Grave, Piotr Bojanowski e Tomas Mikolov: *Bag of tricks for efficient text classification*. arXiv preprint arXiv:1607.01759, 2016. 7, 18, 28

[22] Schütze, Hinrich, Christopher D Manning e Prabhakar Raghavan: *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008. https://nlp.stanford.edu/IR-book/. 7

[23] documentation scikit-learn online: *sklearn.metrics.f1_score*, (acessado em Maio 8, 2021). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score. 10

[24] Davis, Jesse e Mark Goadrich: *The relationship between precision-recall and roc curves*. Em *Proceedings of the 23rd international conference on Machine learning*, páginas 233–240, 2006. 10

[25] Cover, Thomas e Peter Hart: *Nearest neighbor pattern classification.* IEEE transactions on information theory, 13(1):21–27, 1967. 13

[26] Fix, Evelyn: *Discriminatory analysis: nonparametric discrimination, consistency properties.* USAF School of Aviation Medicine, 1951. 13

[27] Levenshtein, Vladimir I: *Binary codes capable of correcting deletions, insertions, and reversals.* Em *Soviet physics doklady*, volume 10, páginas 707–710. Soviet Union, 1966. 14

[28] Yang, Yiming e Xin Liu: *A re-examination of text categorization methods.* Em *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 42–49, 1999. 15

[29] Breiman, Leo, Jerome Friedman, Charles J Stone e Richard A Olshen: *Classification and regression trees.* CRC press, 1984. 16

[30] Quinlan, J. Ross: *Induction of decision trees.* Machine learning, 1(1):81–106, 1986. 16, 29

[31] Quinlan, JR: *Program for machine learning.* C4. 5, 1993. "https://ci.nii.ac.jp/naid/10015645285/en/". 16

[32] Fan, Rong En, Kai Wei Chang, Cho Jui Hsieh, Xiang Rui Wang e Chih Jen Lin: *Liblinear: A library for large linear classification.* the Journal of machine Learning research, 9:1871–1874, 2008. 18, 20

[33] Kleinbaum, David G, K Dietz, M Gail, Mitchel Klein e Mitchell Klein: *Logistic regression.* Springer, 2002. 19, 20

[34] Cortes, Corinna e Vladimir Vapnik: *Support-vector networks.* Machine learning, 20(3):273–297, 1995. 20, 21, 22

[35] George H., John; Kohavi, Ron e Karl Pfleger: *Irrelevant features and the subset selection problem.* Machine Learning Proceedings 1994, Morgan Kaufmann:121–129, 1994. 23

[36] Chen, Jingnian, Houkuan Huang, Shengfeng Tian e Youli Qu: *Feature selection for text classification with naïve bayes.* Expert Systems with Applications, 36(3):5432–5435, 2009. 23, 24

[37] Shannon, Claude E: *A mathematical theory of communication.* The Bell system technical journal, 27(3):379–423, 1948. 24

[38] Salton, Gerard e Christopher Buckley: *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 24(5):513 – 523, 1988, ISSN 0306-4573. http://www.sciencedirect.com/science/article/pii/0306457388900210. 25

[39] Sebastiani, Fabrizio: *Machine learning in automated text categorization.* ACM Comput. Surv., 34(1):1–47, março 2002, ISSN 0360-0300. https://doi.org/10.1145/505282.505283. 28

[40] Lewis, David D e Marc Ringuette: *A comparison of two learning algorithms for text categorization*. Em *Third annual symposium on document analysis and information retrieval*, volume 33, páginas 81–93, 1994. 29

[41] Nacional, Imprensa: *Base de Dados de Publicações do DOU*, (acessado em Maio 11, 2021). `https://www.in.gov.br/acesso-a-informacao/dados-abertos/base-de-dados`. 34

[42] Jabeen, Hafsa: *Stemming and Lemmatization in Python*, 23 Outubro, 2018 (acessado em Maio 8, 2021). `https://www.datacamp.com/community/tutorials/stemming-lemmatization-python`. 37

[43] documentation scikit-learn online: *sklearn.tree.DecisionTreeClassifier*, (acessado em Maio 11, 2021). `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.fit`. 42, 50

[44] Jupyter, Project: *sklearn.tree.DecisionTreeClassifier*, (acessado em Maio 12, 2021). `https://jupyter.org/`. 43