Universidade de Brasília
Faculdade de Tecnologia

Classificação de Publicações em Diários Oficiais Utilizando Aprendizagem de Máquina e Processamento de Linguagem Natural

Igor Furtado Guimarães Estevão

TRABALHO DE GRADUAÇÃO

ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Brasília

2022

Universidade de Brasília
Faculdade de Tecnologia

Classificação de Publicações em Diários Oficiais Utilizando Aprendizagem de Máquina e Processamento de Linguagem Natural

Igor Furtado Guimarães Estevão

Trabalho de Graduação submetido como requisito parcial para obtenção do grau de Engenheiro de Controle e Automação.

Orientador: Prof. Dr. Flávio de Barros Vidal

Brasília
2022

Estevão, Igor Furtado Guimarães.
E79c Classificação de Publicações em Diários Oficiais Utilizando Aprendizagem de Máquina e Processamento de Linguagem Natural / Igor Furtado Guimarães Estevão; orientador Flávio de Barros Vidal. -- Brasília, 2022.
62 p.

Trabalho de Graduação em Engenharia de Controle e Automação -- Universidade de Brasília, 2022.

1. Classificação Multiclasse de Texto. 2. Aprendizagem de Máquina. 3. Processamento de Linguagem Natural. 4. Processo Licitatório de Obras Públicas. I. Vidal, Flávio de Barros, orient. II. Título

Universidade de Brasília
Faculdade de Tecnologia

Classificação de Publicações em Diários Oficiais Utilizando Aprendizagem de Máquina e Processamento de Linguagem Natural

Igor Furtado Guimarães Estevão

Trabalho de Graduação submetido como requisito parcial para obtenção do grau de Engenheiro de Controle e Automação.

Trabalho aprovado. Brasília, 09 de maio de 2022:

Prof. Dr. Flávio de Barros Vidal
UnB/IE/CIC
Orientador

Prof. Dr. Aletéia Patrícia Favacho de Araújo
UnB/IE/CIC
Examinador interno

Prof. Dr. Edison Ishikawa
UnB/IE/CIC
Examinador interno

Brasília
2022

Dedico este trabalho ao meu eu futuro. Espero não lhe levar decepções.

Resumo

De forma a impulsionar o avanço de tecnologias que lidam com a grande quantidade de informação gerada pelo mundo moderno e auxiliar no combate à corrupção, o presente trabalho buscou desenvolver uma metodologia de classificação de publicações relacionadas ao processo licitatório de obras públicas, disponíveis no Diário Oficial da União, quanto ao tipo de publicação, sendo esse um problema de classificação multiclasse de texto. Para isso, foi reunido um conjunto de dados com 4.181.390 publicações, o qual somado ao uso de técnicas de aprendizagem de máquina e processamento de linguagem natural, possibilitaram a simulação de processos de classificação reais considerando diferentes cenários. Foram testados 14 modelos lineares clássicos diferentes na resolução do problema, os quais tiveram seus desempenhos medidos e comparados por meio do cálculo de métricas estatísticas derivadas da matriz de confusão multiclasse. Os resultados obtidos foram capazes de manter o nível de qualidade do que seria o trabalho de um ser humano capacitado, tendo o modelo *LinearSVM-L2* atingido um F1-*score* de 97.88% em um dos cenários, mostrando que as técnicas utilizadas são muito eficazes na resolução de problemas dessa natureza e abrindo caminho para a resolução de problemas mais complexos.

Palavras-chave: Classificação Multiclasse de Texto. Aprendizagem de Máquina. Processamento de Linguagem Natural. Processo Licitatório de Obras Públicas.

Abstract

In order to boost the advancement of technologies that deal with the large amount of information generated by the modern world and assist in the fight against corruption, the present work sought to develop a methodology for classifying publications related to the bidding process of constructions, avaiable at Diário Oficial da União, regarding the type of publication, this being a multiclass text classification problem. To do that, a dataset with 4.181.390 publications was gathered, which, together with machine learning and natural language processing techniques, allowed the simulation of open-world classification processes considering different scenarios. Fourteen different classical linear models were tested, which had their performances measured and compared by calculating statistical metrics derived from the multiclass confusion matrix. The results obtained were able to maintain the quality level of what would be the work of a trained human being, with the *LinearSVM-L2* model reaching an F1-*score* of 97.88% in one of the scenarios, showing that the techniques used are very effective in solving problems of this nature and opening the path to solving more complex problems.

Keywords: Multiclass Text Classification. Machine Learning. Natural Language Processing. Public Works Bidding Process.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under the Curve* 37
CADE	Conselho Administrativo de Defesa Econômica 13
CRF	*Conditional Random Field* 40
DOU	Diário Oficial da União 14
DTC	*Decion Tree Classifier* 40
DV	*Deep Vacuity* 12
FN	Falsos Negativos 34
FP	Falsos Positivos 34
HTML	*HyperText Markup Language* 24
INC	Instituto Nacional de Criminalística 38
JSON	*JavaScript Object Notation* 42
kNN	*k-Nearest Neighbors* 25
LR	*Logistic Regression* 40
NBC	*Naive Bayes Classifier* 40
NLP	*Natural Language Processing* 13
PF	Polícia Federal 15
ROC	*Receiver Operating Characteristic* 37
SVM	*Support Vector Machine* 40
TFP	Taxa de Falsos Positivos 37
TVP	Taxa de Verdadeiros Positivos 37
UnB	Universidade de Brasília 38
VN	Verdadeiros Negativos 33
VP	Verdadeiros Positivos 33
XML	*Extensible Markup Language* 43

Sumário

1 Introdução

Neste capítulo são criadas as premissas necessárias ao entendimento do que será realizado neste trabalho, sendo isso feito por meio da exposição das ideias motivadoras por trás de sua idealização, da descrição do problema a ser resolvido, e das definições da hipótese de pesquisa e dos objetivos a serem alcançados. Ao final do capítulo também são mostradas a estrutura e organização dos capítulos seguintes.

1.1 Motivação

É importante dizer que as motivações por trás da idealização deste trabalho estão diretamente ligadas às do *Deep Vacuity* (DV), que é um projeto maior no qual este trabalho está inserido. O DV tem como objetivo o desenvolvimento de um método de identificação de formação de cartel em licitações de obras públicas brasileiras por meio do uso de técnicas de aprendizado de máquina e inteligência artificial. Mais informações sobre esse projeto e de como ele se relaciona com este trabalho podem ser encontradas na Seção 3.1. Para melhor entendimento dessas ideias motivadoras, as mesmas foram divididas por campo de influência e expostas nas próximas duas subseções.

1.1.1 No Âmbito Social

Vem se tornando cada vez mais desafiador lidar com a enorme quantidade de informação gerada diariamente pelo mundo moderno e globalizado, sendo necessário cada vez mais apelar para métodos computacionais de forma a simplificar a resolução de problemas e facilitar a vida das pessoas. Isso não é diferente no caso do combate à corrupção sistêmica, que é um problema em alta no Brasil e que exige que toda a informação produzida pelas ações do governo seja constantemente monitorada (BOCHENEK; PEREIRA, 2018). Além disso, devido à ocorrência de grandes investigações policiais nos últimos anos, somado à forma como a mídia vem cobrindo essas operações, houve um aumento significativo na demanda pelo combate à corrupção por parte da população, o que causou também um aumento na pressão em cima dos órgãos públicos responsáveis (AZEVEDO, 2010). A Operação Lava Jato foi um dos grandes agravantes dessa situação, pois além de ser considerada por muitos como a maior operação anticorrupção da história do Brasil, foi uma das causas da crise política e econômica que se instaurou em 2014, gerando insatisfação popular e protestos contra o governo por todo o país (CIOCCARI, 2015).

Isso tudo é evidente no combate à fraude em licitações de obras públicas, que é um dos principais desafios que vem sendo enfrentados pela sociedade na guerra contra a corrupção. Essas fraudes envolvem grandes quantias em dinheiro e alta complexidade técnica e burocrática, além de darem muita margem para práticas maliciosas (CADE, 2019). Isso é discutido no Guia de Combate a Cartéis em Licitação, que foi elaborado pelo Conselho Administrativo de Defesa Econômica (CADE). Esses processos produzem uma quantidade exorbitante de informação, fazendo com que seja extremamente difícil para os órgãos públicos encarregados supervisionar tudo e garantir que nada de ilegal esteja ocorrendo. Além disso, muitas vezes há membros internos envolvidos nos esquemas, fazendo com que muita coisa passe batido e grandes somas em dinheiro sejam desviadas todo ano (CASTRO, 2010). Quem mais sofre com tudo isso é o povo brasileiro, visto que os recursos desviados deveriam estar sendo usados em prol de melhorar a condição de vida das pessoas.

Tendo tudo isso em mente, qualquer melhoria ou avanço em técnicas ou metodologias que sejam úteis na fiscalização desses processos ou ajudem a lidar com a grande quantidade de informação gerada por eles, seria de grande ajuda no combate à corrupção e pode ser de grande benefício para a sociedade.

1.1.2 No Âmbito Científico e Acadêmico

Os recentes avanços na área de processamento de linguagem natural (NLP, que é acrônimo em inglês para *Natural Language Processing*) tem impulsionado cada vez mais o surgimento de novas pesquisas, levando à solução de novos problemas todos os dias. Isso gera uma bola de neve, fazendo com que a área avance cada vez mais, e mais pesquisas sejam feitas. Esse crescimento exponencial possibilita a solução de problemas cada vez mais complexos, impulsionando o avanço em diversas áreas do conhecimento e causando um impacto positivo direto na vida das pessoas (CHOWDHARY, 2020).

Lidar com a enorme quantidade de informação gerada pelo mundo moderno e globalizado é um dos principais desafios enfrentados pelo NLP, o qual tem se tornado cada vez mais difícil, visto que a taxa de geração de dados só tende a crescer e o nível de complexidade estrutural dos mesmos só a aumentar (GANTZ; REINSEL, 2012). Considerando o investimento em pesquisa que ainda será feito de forma com que o avanço tecnológico consiga acompanhar essa crescente demanda, muitas oportunidades valiosas serão criadas, fazendo com que avanços expressivos na área sejam muito bem recompensados.

Tendo tudo isso em mente, pesquisas acadêmicas que produzam avanços expressivos na forma como armazenamos, tratamos e analisamos os nossos dados pode ser uma grande conquista para a ciência.

1.2 Descrição do Problema

Este trabalho lida com algo semelhante aos problemas precursores das ideias motivadoras mencionadas na seção anterior, tendo o problema aqui tratado origem na grande quantidade de informação gerada pelos processos licitatórios de obras públicas e nas irregularidades presentes nos mesmos. É previsto por lei que todos os detalhes referentes a cada uma das etapas desses processos sejam publicados na Seção 3 do Diário Oficial da União (DOU), e qualquer pessoa tenha acesso a essas informações. Assim sendo, a Imprensa Nacional disponibiliza um banco de dados digital com todos os textos de publicações feitas a partir do ano de 2002 [1] (BRASIL, 1993). Mais informações a respeito do DOU e sua relação com o que será feito aqui podem ser encontradas na Seção 2.2. A Figura 1 mostra parte da primeira página referente à Seção 3 do jornal publicado no dia 08 de abril de 2022. É evidenciado no sumário que essa edição possui 304 páginas, e considerando que a maior parte das páginas tem mais de 10 publicações, é possível ter uma ideia da enorme quantidade de informação disponibilizada diariamente. Isso leva ao pensamento de que seja bem provável que uma boa parte desses dados não seja monitorada.

ISSN 1677-7069

DIÁRIO OFICIAL DA UNIÃO

REPÚBLICA FEDERATIVA DO BRASIL • IMPRENSA NACIONAL

Ano CLX Nº 68 — Brasília - DF, sexta-feira, 8 de abril de 2022 — SEÇÃO 3

Sumário

Esta edição é composta de 304 páginas

Presidência da República

SECRETARIA-GERAL

SUBCHEFIA PARA ASSUNTOS JURÍDICOS

EXTRATO DE TERMO ADITIVO AO ACORDO DE COOPERAÇÃO TÉCNICA

TERMO ADITIVO Nº 01 AO ACORDO DE COOPERAÇÃO TÉCNICA ENTRE: A Subchefia para Assuntos Jurídicos da Secretaria - Geral da Presidência da República, CNPJ nº 00.394.411/0001-09, e a Pontifícia Universidade Católica do Paraná. ESPÉCIE: Termo Aditivo ao Acordo de Cooperação Técnica nº 02/2019 (Processo nº 00025.0000176/2019-00). OBJETO: Inclusão da Cláusula da Lei Geral de Proteção de Dados. DATA DE ASSINATURA: 05/04/2022, Pedro Cesar Nunes F. M. de Souza, Subchefe para Assuntos Jurídicos da Secretaria - Geral da Presidência da República, e Rogério Renato Mateucci, Reitor da Pontifícia Universidade Católica do Paraná.

EXTRATO DE TERMO ADITIVO Nº 3/2022 - UASG 110099 - SAD/SP/AGU

Número do Contrato: 11/2019.
Nº Processo: 00589.000332/2019-49.
Pregão. Nº 1/2019. Contratante: SUPERINTENDENCIA ADMINISTRACAO EM SAO PAULO. Contratado: 21.862.782/0001-48 - MV CLEAN SERVICOS TECNICOS E CONSERVACAO LTDA. Objeto: 1.1. O objeto do presente instrumento é:
1.1.1. Prorrogar o prazo da vigência do contrato nº 11/2019, por 12 (doze) meses, contemplando-se, nesta ocasião, o período de 26/04/2022 a 25/04/2023, nos termos do art. 57, ii, da lei n.º 8.666, de 1993.
1.1.2. Adequar o contrato às alterações trazidas pela instrução normativa seges/me nº 53, de 8 de julho de 2020.. Vigência: 26/04/2022 a 25/04/2023. Valor Total Atualizado do Contrato: R$ 45.671,88. Data de Assinatura: 01/04/2022.

(COMPRASNET 4.0 - 01/04/2022).

SUPERINTENDÊNCIA DE ADMINISTRAÇÃO NO DISTRITO FEDERAL

AVISO DE REABERTURA DE PRAZO
PREGÃO Nº 3/2022

Comunicamos a reabertura de prazo da licitação supracitada, processo Nº 00590000627202109. , publicada no D.O.U de 09/02/2022 . Objeto: Pregão Eletrônico - Contratação de empresa especializada na prestação de serviços nas áreas de edição de mídias audiovisuais e design gráfico, mediante cessão de mão de obra exclusiva, para atender necessidades da Assessoria de Comunicação Social da Advocacia-Geral da União em Brasília/DF, conforme condições, quantidades e exigências estabelecidas no Edital e seus anexos. Novo Edital: 08/04/2022 das 08h00 às 12h00 e de14h00 às 17h59. Endereço: Sig Quadra 06 Lote 800 Sig - BRASILIA - DFEntrega das Propostas: a partir de 08/04/2022 às 08h00 no site www.comprasnet.gov.br. Abertura das Propostas: 25/04/2022, às 10h00 no site www.comprasnet.gov.br.

RODRIGO JORG PFEILSTICKER
Superintendente

(SIDEC - 07/04/2022) 110161-00001-2022NE000096

GABINETE DE SEGURANÇA INSTITUCIONAL

AGÊNCIA BRASILEIRA DE INTELIGÊNCIA

AVISO DE SUSPENSÃO
PREGÃO Nº 9/2022

Comunicamos a suspensão da licitação supracitada, publicada no D.O.U em 29/03/2022 . Objeto: Pregão Eletrônico - Contratação de empresa especializada para execução de serviços de Agente de Integração de Estágio em favor da Agência de Brasileira de Inteligência (ABIN), responsável por todo o processo de contratação de estagiários de nível superior, nas modalidades de graduação e pós graduação, desde a seleção até o desligamento do estagiário, incluindo a intermediação e o pagamento de seguro contra acidentes pessoais.

ESLONY BISPO DOS SANTOS
Pregoeiro

Figura 1 – Parte da primeira página referente à Seção 3 do jornal publicado em 08 de abril de 2022.

Fonte: (BRASIL, 2022).

[1] Esse banco de dados pode ser acessado pelo endereço: https://www.in.gov.br/acesso-a-informacao/dados-abertos/base-de-dados

Vários órgãos públicos ficam encarregados pela fiscalização desses processos, sendo a Polícia Federal (PF) um deles, e muitos recursos precisam ser direcionados para acompanhar de perto essa grande quantidade de informação e garantir que tudo esteja acontecendo dentro do que a lei permite. Dentro deste contexto, este trabalho procura auxiliar no trabalho da PF por meio do desenvolvimento de uma metodologia de classificação de publicações relacionadas ao processo licitatório de obras públicas quanto ao tipo de publicação, sendo isso feito por meio do uso de técnicas de aprendizado de máquina e de processamento de linguagem natural. A Figura 2 mostra um esquema ilustrativo desse problema, no qual a interrogação representa o método de classificação a ser desenvolvido. O seu funcionamento irá consistir no recebimento de um texto de publicação como entrada e da impressão na saída de um rótulo que indica o tipo de publicação a qual aquele texto representa.

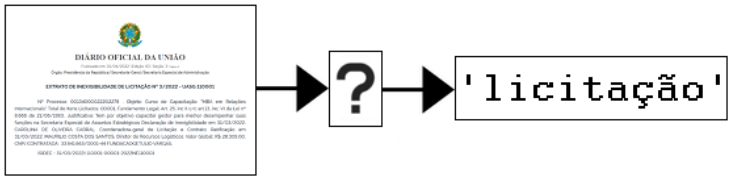

Figura 2 – Esquema ilustrativo do problema a ser resolvido.

Figura produzida pelo autor.

Para que um problema como esse possa ser resolvido por meio da utilização das técnicas propostas, o primeiro passo é obter um grande conjunto de textos de publicações de forma com que essa amostra represente fielmente o todo. Dessas publicações podem ser extraídos dados relevantes, e a partir desses dados pode ser feito um estudo sistemático, identificando fatores determinantes na diferenciação entre importantes tipos de publicações. Com isso, as publicações podem ser agrupadas em classes, permitindo a criação de um conjunto de dados que possa ser utilizado em aprendizado de máquina. Em seguida, técnicas de processamento de linguagem natural podem ser utilizadas para transformar os dados em algo mais fácil de ser trabalhado. A partir daí podem ser considerados possíveis caminhos para o desenvolvimento da metodologia de classificação, podendo se fazer testes com métodos já consolidados ou tentativas de criação de novos. Por fim, o método de melhor desempenho pode ser apontado por meio de uma análise comparativa entre os resultados.

1.3 Hipótese de Pesquisa

A hipótese de pesquisa deste trabalho se baseia em verificar se o uso de aprendizado de máquina em conjunto com técnicas de processamento de linguagem natural é suficiente para classificar publicações informativas relacionadas ao processo licitatório de obras públicas brasileiras, disponíveis na Seção 3 do DOU, quanto ao tipo de publicação, e obter resultados equiparáveis aos que seriam os do trabalho de um ser humano capacitado.

1.4 Objetivos

Os objetivos do trabalho foram divididos em objetivo primário e objetivos secundários, sendo os mesmos definidos e listados a seguir:

1.4.1 Objetivo Primário

O objetivo primário ou principal é criar uma metodologia de classificação de publicações informativas relacionadas ao processo licitatório de obras públicas brasileiras, disponíveis na seção 3 do DOU, quanto ao tipo de publicação. Essa metodologia será desenvolvida utilizando aprendizado de máquina em conjunto com técnicas de processamento de linguagem natural e precisa manter o nível de qualidade do que seria o trabalho feito por um especialista.

1.4.2 Objetivos Secundários

Abaixo são listados os objetivos secundários deste trabalho:

- Criar um novo conjunto de dados com textos de publicações divididos em classes para uso em possíveis trabalhos futuros relacionados;
- Avaliar o desempenho de diferentes modelos de classificação de texto na resolução do problema definido utilizando diferentes configurações e parâmetros;
- Comparar o impacto causado pelo uso de diferentes procedimentos e técnicas de pré-processamento nos resultados finais do trabalho;
- Contribuir para o avanço do projeto DV na busca pelos seus objetivos.

1.5 Organização do Manuscrito

O conteúdo remanescente deste relatório é estruturado e organizado conforme mostrado a seguir: o Capítulo 2 descreve os tópicos julgados necessários ao entendimento do trabalho; o Capítulo 3 apresenta o projeto DV e mostra conexões entre o que será feito aqui e outros trabalhos relevantes da área; o Capítulo 4 descreve as etapas do processo de desenvolvimento do trabalho de forma detalhada; o Capítulo 5 apresenta os resultados finais do trabalho enquanto levanta algumas dificuldades encontradas, e discute aspectos ligados à tomada de decisão; e, por fim, o Capítulo 6 compara os resultados obtidos com os esperados inicialmente, e aponta direções para o desenvolvimento de trabalhos futuros.

2 Fundamentação Teórica

Este capítulo reúne informações a respeito e descreve todos os tópicos considerados necessários ao entendimento das próximas etapas deste trabalho, servindo como embasamento teórico para o seu desenvolvimento e garantindo a compreensão do leitor. Assim, são destacados os assuntos relacionados à classificação de texto e as técnicas de processamento de linguagem natural a serem utilizadas, visto que formam o núcleo teórico principal do trabalho. Além disso, são também descritos aspectos ligados aos processos licitatórios de obras públicas, ao Diário Oficial da União, às técnicas de aprendizado de máquina e aos métodos a serem utilizados na avaliação dos resultados.

2.1 Processo Licitatório de Obras Públicas

A licitação, que é regulamentada pela Lei de número 8.666/1993, é a forma prevista pela Constituição Federal para que sejam realizadas contratações de serviços ou compras de produtos pela Administração Pública. Essa lei também prevê que seja obrigatória a publicação no DOU de todos os atos oficiais relativos a processos licitatórios de compras ou serviços, quando se tratar de licitação que tenha qualquer envolvimento de algum órgão ou entidade da Administração Pública Federal (BRASIL, 1993).

Devido à natureza extremamente burocrática dos processos licitatórios de obras públicas, uma grande quantidade de informação é produzida na forma de atos oficiais a todo momento, bombardeando o DOU com centenas de publicações todos os dias. Esses atos são oriundos de avanços nos processos licitatórios e estão diretamente ligados à etapa em que os mesmos se encontram. As próximas subseções descrevem essas etapas e mostram como elas se relacionam com os atos em questão, sendo também listadas as principais atividades fraudulentas encontradas nesses processos, e mencionada uma nova lei que pode alterar a forma como esses atos serão publicados em um futuro próximo.

2.1.1 Etapas do Processo Licitatório

Os processos licitatórios podem seguir diferentes procedimentos em sua realização de acordo com o que é especificado pela modalidade de licitação escolhida. Existem cinco modalidades convencionais: concorrência, tomada de preços, convite, concurso e leilão, os quais foram também definidas pela Lei 8.666/1993 (BRASIL, 1993), e o pregão, que é uma modalidade um pouco mais recente criada pela Lei 10.520/2002 (BRASIL, 2002) e que vem sendo priorizada devido aos benefícios que traz à Administração Pública (MOREIRA; GUIMARÃES, 2012). A existência dessas modalidades visa proporcionar um certo grau de

flexibilidade, possibilitando o ajuste do processo à situação em questão, visando sempre obter o melhor resultado possível a partir da consideração dos preços, prazos e qualidade dos serviços e/ou produtos a serem adquiridos. Somente algumas delas são utilizadas em licitações de obras públicas: a **concorrência**, a **tomada de preços** e o **convite**. Além dessas modalidades, esses processos são divididos em duas fases: a interna e a externa. A primeira fase delas, que também é chamada de preparatória, envolve a burocracia interna dos órgãos governamentais envolvidos, sendo também onde é avaliada a proposta de licitação, a modalidade de licitação a ser utilizada é definida e o edital é elaborado. A segunda é a partir de onde os detalhes da licitação vem a público, envolvendo a participação de empresas interessadas e abrindo margem para atividades fraudulentas. Esta fase é a origem das informação que compõe as publicações feitas no DOU, e será o objeto de interesse deste trabalho. As etapas da fase externa do processo licitatório são descritas a seguir:

- **Abertura:** etapa correspondente ao momento em que a licitação torna-se pública, constituindo da liberação do edital ou ato convocatório e permitindo que interessados se prepararem para a possível participação;
- **Habilitação:** etapa na qual é verificado se os interessados preenchem os requisitos impostos e dispõe dos documentos exigidos, garantindo que estejam habilitados a participar das próximas etapas;
- **Julgamento:** etapa onde as propostas são avaliados e julgadas de acordo com os critérios definidos pelo edital de convocação, sendo em seguida organizadas e classificadas;
- **Homologação:** nesta etapa o resultado da licitação é reconhecido pelos órgãos responsáveis, sendo ratificado todo o procedimento e conferido aos atos licitatórios aprovação para que produzam os efeitos jurídicos necessários;
- **Adjudicação:** é a última etapa do processo, onde é determinado o vencedor da licitação e lhe é conferido o objeto em questão.

2.1.2 Nova Lei de Licitações e Contratos Administrativos

Uma nova lei para a regularização do processo licitatório de obras públicas foi sancionada em 1 de abril de 2021, a Lei de Licitações e Contratos Administrativos, de número 14.133/2021. Ela substitui as Leis 8.666/1993, 10.520/2002 e 12.462/2001, que em conjunto ficam por conta do mesmo trabalho no momento (BRASIL, 2021).

Apesar da substituição, as leis antigas ainda serão válidas por dois anos a partir da data de sancionamento da nova lei, estando este trabalho ainda sob suas vigências. É esperado que essa substituição cause mudanças na forma como atos referentes ao processo licitatório de obras públicas sejam publicados, causando efeito direto nos resultados deste trabalho. Não há ainda nenhuma informação disponível sobre essas possíveis mudanças.

2.2 Diário Oficial da União

O Diário Oficial da União, de sigla DOU, é um jornal governamental que funciona como canal oficial de comunicação e documentação do Governo Federal, sendo responsável por tornar público qualquer assunto que seja relevante e/ou de interesse da população. Esse jornal existe desde 1862, estando disponível apenas no formato impresso até 2001, quando passou a também a oferecer edições eletrônicas. Em dezembro de 2017, começou a ser disponibilizado exclusivamente no formato digital, ficando a responsabilidade de distribuição por conta da Imprensa Nacional e o acesso pelo endereço `www.in.gov.br`. As próximas subseções reúnem informações relevantes acerca de como é publicado no DOU e sobre o conteúdo dessas publicações, sendo dada uma importância a mais às relacionadas aos processos licitatórios de obras públicas.

2.2.1 Organização e Conteúdo das Publicações

Por meio da exigência da publicação de certos tipos de atos o DOU busca tornar públicas e transparentes as ações da Administração Pública e de empresas associadas em tudo o que diz respeito ao interesse da sociedade. De forma geral, os principais tipos de atos que devem ser publicados são: atos ou instrumentos normativos que sejam exigidos como obrigação legal; atos de interesse de servidores dos Poderes Legislativo e Judiciário, do Ministério Público da União, civis e militares da União e qualquer outro colaborador da Administração Pública; atos relacionados a licitações e contratos públicos; e atos de entidades públicas que dizem respeito às informações de cunho econômico e financeiro (E-DOU, 2021).

A publicação no DOU é organizada por meio da divisão dos atos a serem publicados em três seções, a qual é feita com base em seus tipos conforme mostrado a seguir:

- **Seção 1:** reservada aos atos normativos de interesse geral dos poderes da União, sendo publicados leis, decretos, resoluções, instruções normativas, portarias, tratados, etc;
- **Seção 2:** reservada aos atos de pessoal relativos aos servidores da Administração Pública Federal, sendo publicados decisões, decretos de pessoal, despachos, etc;
- **Seção 3:** reservada aos atos decorrentes das contratações públicas e outros de particulares, sendo publicados avisos de licitação, acordos, autorizações de compra, contratos, convênios, comunicados, notificações de concursos públicos, intimações, etc.

Considerando que o objeto de interesse são as publicações relacionadas aos processos licitatórios de obras públicas, serão somente utilizadas neste trabalho as feitas na Seção 3, visto que a publicação desses tipos de atos está restringida à seção em questão.

2.2.2 Publicações Relacionadas ao Processo Licitatório de Obras Públicas

As publicações no DOU são regidas pelo Decreto de Número 9.215/2018, sendo estabelecido que a publicação de editais, licitações, contratos, convênios e aditivos deve ser feita na forma de extratos, que são resumos que se restringem somente aos elementos necessários à identificação do ato em questão (BRASIL, 2017).

A Figura 3 mostra um exemplo de publicação feita na Seção 3 do DOU, representando um extrato de inexigibilidade de licitação. Esse tipo de ato firma uma contratação sem competição em razão da inviabilidade da mesma ou da desnecessidade do procedimento licitatório. Assim é possível perceber o quão resumida a informação está, tendo o mínimo de palavras desnecessárias e informações condensadas em sequência.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 01/04/2022 | Edição: 63 | Seção: 3 | Página: 1

Órgão: Presidência da República/Secretaria-Geral/Secretaria Especial de Administração

EXTRATO DE INEXIGIBILIDADE DE LICITAÇÃO Nº 3/2022 - UASG 110001

Nº Processo: 00134000022202278 . Objeto: Curso de Capacitação: "MBA em Relações Internacionais". Total de Itens Licitados: 00001. Fundamento Legal: Art. 25, inc II c/c art.13, inc. VI da Lei nº 8.666 de 21/06/1993.. Justificativa: Tem por objetivo capacitar gestor para melhor desempenhar suas funções na Secretaria Especial de Assuntos Estratégicos Declaração de Inexigibilidade em 31/03/2022. CAROLINA DE OLIVEIRA CABRAL. Coordenadora-geral de Licitação e Contrato. Ratificação em 31/03/2022. MAURILIO COSTA DOS SANTOS. Diretor de Recursos Logísticos. Valor Global: R$ 28.309,00. CNPJ CONTRATADA : 33.641.663/0001-44 FUNDACAOGETULIO VARGAS.

(SIDEC - 31/03/2022) 110001-00001-2022NE110001

Figura 3 – Exemplo de publicação feita na Seção 3 do Diário Oficial da União.

Fonte: adaptado de (BRASIL, 2022).

Todas as publicações relacionadas a processos licitatórios de obras públicas seguem uma estrutura semelhante à do exemplo acima, tendo principalmente: um título que remete ao tipo de ato que foi publicado; datas; números de identificação do ato, da publicação e de processos associados; informações numéricas se referindo a pessoas e/ou empresas (CPF e/ou CNPJ); informações numéricas referentes a valores; nomes de pessoas envolvidas no processo; informações numéricas referentes a leis; e algumas palavras explicando o objeto em questão e fazendo ligação entre todas essas informações.

Os tipos de publicações mais importantes com origem nos processos licitatórios são as licitações, que consistem em editais e outras informações a respeito do processo; contratos, que são a efetivação do objeto de um processo licitatório; convênios, que são a formação de uma parceria entre órgãos envolvidos em uma mesma licitação; e termos aditivos, que são modificações ou extensões de contratos já feitos.

2.3 Processamento de Linguagem Natural

O processamento de linguagem natural, também chamado de NLP, que é sigla do inglês para *Natural Language Processing*, é um campo de estudo que se interessa pelas interações entre a linguagem humana e os computadores, sendo resultante da intersecção entre ciência da computação, inteligência artificial e linguística (VAJJALA et al., 2020).

A sua principal preocupação é com o desenvolvimento de sistemas computacionais que sejam capazes de entender e analisar a nossa linguagem, permitindo que os mesmos sejam utilizados para impulsionar avanços em diversas áreas do conhecimento. A Figura 4 mostra algumas aplicações expressivas presentes em nossas vidas, e que possuem alguma técnica de NLP como componente chave de seu desenvolvimento.

Figura 4 – Exemplos de aplicações expressivas de NLP presentes em nossas vidas.

Fonte: adaptado de Vajjala et al. (2020).

O alcance do NLP vem aumentando muito nos últimos anos, já tendo aplicações por todos os lugares. Alguns de seus principais campos de atuação são listados e descritos a seguir, sendo também mencionadas algumas aplicações interessantes (GOLDBERG, 2017):

- **Criação de modelos de linguagem (*language modeling*):** consiste na predição dos próximos elementos de um texto baseada no histórico de elementos anteriores. Os resultados dessa área podem ser úteis na geração de conversas automatizadas, reconhecimento de fala e escrita, geração de textos, entre outras aplicações;
- **Classificação de texto (*text classification*):** consiste em categorizar textos a partir de rótulos predefinidos. Mais detalhes a respeito desta área serão vistos na Seção 2.4;
- **Extração de informação (*information extraction*):** consiste na busca por informação relevante dentro de um texto. Pode ser útil na busca por nomes de pessoas mencionadas em *posts* de redes sociais e na busca por eventos em *emails*;

- **Recuperação de informação (*information retrieval*):** consiste em encontrar um conjunto de dados relacionado a um tipo de informação desejada dentro de uma grande quantidade de dados. O principal exemplo de aplicação desta área é o *Google Search* [1];
- **Análise de sentimentos (*sentiment analysis*):** consiste em interpretar e classificar possíveis informações subjetivas a partir de textos. É muito utilizado em aplicações de interação com usuários em redes sociais e em sistemas de atendimento automático;
- **Agentes de conversação (*conversational agent*):** consiste na criação de sistemas capazes de reproduzir diálogo humano. A Siri e a Alexa [2] são exemplos desses sistemas;
- **Criação de resumo de texto (*text summarization*):** consiste na criação de resumos de textos enquanto são mantidas informações relevantes e o significado do mesmo;
- **Tradução (*machine translation*):** consiste na tradução de um texto de uma linguagem para outra. O principal exemplo de aplicação desta área é o *Google Translate* [3];
- **Sistemas de resposta automática (*question answering*):** consiste no treinamento de um modelo para responder perguntas a partir de um texto de referência.

Como demonstrado acima, a área de atuação de NLP é bem vasta. Por isso, o foco de uma descrição mais a fundo se limitará aos aspectos relacionados à classificação de texto, que além de ser de longe seu principal campo de atuação, faz uso de grande parte de suas técnicas e metodologias. Essa descrição será feita ao longo da Seção 2.4.

2.4 Classificação de Texto

A classificação de texto é uma tarefa que consiste na categorização de texto em rótulos predeterminados, podendo ser feita em nível de documentos, frases ou palavras, e podendo envolver somente duas classes (binária) ou múltiplas classes. Os modelos utilizados na resolução desse problema recebem, normalmente, como entrada informação de texto bruta, sendo as mesmas sequências de texto em documentos como $D = (X_1, X_2, ..., X_n)$, onde X_i pode se referir a qualquer ponto no texto que delimite um segmento. Cada um desses pontos é então rotulado com um valor de classe oriundo de um conjunto de k diferentes valores discretos predeterminados (KOWSARI et al., 2019).

[1] Essa aplicação pode ser acessada pelo endereço: https://www.google.com.br/
[2] Mais informações nos endereços: https://www.apple.com/siri/ https://alexa.amazon.com/
[3] Essa aplicação pode ser acessada pelo endereço: https://translate.google.com/

As próximas subseções irão descrever as técnicas ou metodologias envolvidas no *pipeline* mais comum utilizado na resolução de problemas de classificação de texto, conforme ilustrado na Figura 5.

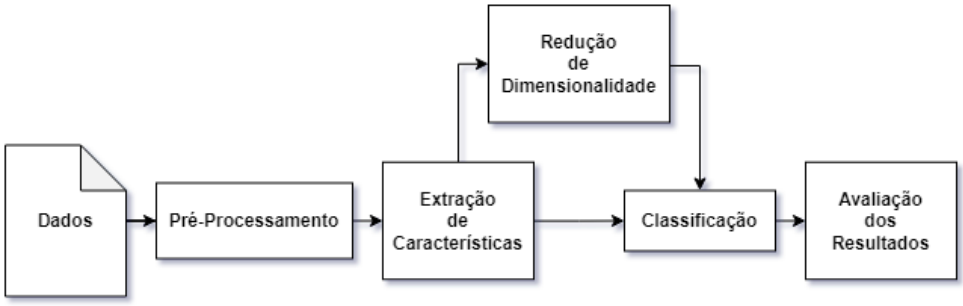

Figura 5 – Diagrama representando o fluxo de etapas de um processo de classificação de texto.

Fonte: adaptado de (KOWSARI et al., 2019).

2.4.1 Pré-Processamento de Texto

Muitos algoritmos que lidam com problemas de aprendizado de máquina sofrem queda de desempenho quando lidam com a presença de ruído e/ou informações irrelevantes à solução no conjunto de dados com o qual está trabalhando, e isso não é diferente no âmbito textual (VIJAYARANI; ILAMATHI; NITHYA et al., 2015). A seguir são listadas e brevemente descritas todas as técnicas de pré-processamento de texto consideradas relevantes ao desenvolvimento deste trabalho.

Tokenização (*Tokenization*)

A tokenização é uma técnica que consiste na divisão de uma sequência textual em palavras, frases, letras, ou outros elementos de interesse chamados *tokens*. Isso é feito com o objetivo de facilitar a análise do texto e atender a requisitos de alguns algoritmos de NLP (KOWSARI et al., 2019).

Palavras de Parada (*Stop Words*)

As *stop words* são basicamente palavras que não agregam valor nenhum ao texto no ponto de vista dos algoritmos de NLP. Essas palavras costumam em sua grande maioria serem artigos e preposições, dependendo diretamente do idioma no qual o texto está escrito, e removê-las do texto é quase sempre a melhor opção.

Capitalização (*Capitalization*)

A uniformização da capitalização no texto é de extrema importância em problemas de classificação textual, visto que não seguir uma regra pode causar inconsistência na interpretação e redução drástica no desempenho do classificador. A abordagem mais comum para lidar com essas situações é a minusculização de todo o texto. Essa técnica precisa ser

utilizada com cuidado, pois depende muito do problema em questão, visto que utilizada no lugar errado pode causar impossibilidade de interpretação de algumas palavras.

Remoção de Ruído (*Noise Removal*)

Consiste no processo geral de remoção de qualquer tipo de informação irrelevante ao problema, estando associada com várias das outras técnicas mencionadas nessa seção. Símbolos, pontuação, marcações HTML e resíduos de estruturas de texto são exemplos de ruido passíveis de serem removidos.

Stemização (*Stemming*)

É uma técnica que visa transformar diferentes formas de palavras que tem a mesma semântica em um único tipo de *feature* ou característica, sendo isso feito por meio da redução de palavras à sua forma raiz, ou *stem*.

Lematização (*Lemmatization*)

Lematização no contexto da classificação de texto é uma técnica de NLP que busca reduzir palavras à sua forma básica (lemas), de forma a agrupar as que se assemelham e com isso reduzir a informação dispensável do texto. Essa técnica também leva em consideração o contexto da palavra de forma a resolver problemas semânticos, como a desambiguação, o que significa que ela pode discriminar entre palavras idênticas que têm significados diferentes, dependendo do contexto específico.

2.4.2 Extração de Características

No âmbito da classificação de texto, a extração de características ou *features* se refere a um conjunto de técnicas aplicadas ao texto que o modifica de forma a estruturá-los como informação, e fazer com que um modelo de aprendizagem de máquina consiga entendê-lo. As técnicas mais utilizadas em problemas de classificação de texto para a extração de características são: TF-IDF, TF, *Word2Vec* e *Global Vectors for Word Representation* (GloVe).

O TF-IDF (acrônimo do inglês *Term Frequency-Inverse Document Frequency*), que é a técnica a ser utilizada neste trabalho, baseia-se no cálculo da frequência de ocorrência de palavras no texto, sendo que TF se refere a frequência de termos (*Term Frequency*) e IDF se refere a frequência inversa de documentos (*Inverse Document Frequency*). Essa técnica atribui um peso a cada palavra de acordo com a quantidade de vezes que a mesma aparece no texto/documento, sendo dado um peso alto tanto para palavras que aparecem muito quando para as que aparecem pouco. A ideia por trás disso é tentar minimizar o efeito das palavras que aparecem muito no texto mas não carregam significado, como artigos, preposições e etc. (RAMOS et al., 2003). A Equação 2.1 descreve matematicamente essa técnica, onde N é o número de documentos e $df(t)$ é o número de documentos em que a palavra t aparece (KOWSARI et al., 2019).

$$W(d,t) = TF(d,t) * log(N/df(t)) \tag{2.1}$$

2.4.3 Redução de Dimensionalidade

A redução de dimensionalidade é um método que consiste na redução do número de variáveis de entrada do conjunto de dados (dimensionalidade) de forma a reduzir o processamento computacional requerido em seu processamento, reduzindo tanto o consumo de tempo quanto a ocupação de memória. As principais técnicas utilizadas na redução de dimensionalidade de dados são: Análise de Componentes (*Component Analysis*), Análise Linear do Discriminante (*Linear Discriminant Analysis*), Fatoração de Matriz Não Negativa e Projeção Aleatória (*Random Projection*) (VAN DER MAATEN; POSTMA; VAN DEN HERIK et al., 2009).

Essas técnicas foram citadas de forma a caracterizar completamente o *pipeline* frequentemente utilizado em problemas de classificação de texto, mas não serão utilizadas neste trabalho, visto que processamento computacional não é um problema, e informações importantes podem ser perdidas com o seu uso.

2.4.4 Métodos de Classificação

A etapa mais importante na resolução de um problema de classificação de texto é a escolha do método que melhor se adéqua à situação em questão. Os métodos de classificação de texto a serem utilizados neste trabalho são descritos um a um nas próximas subseções.

2.4.4.1 kNN (k-Nearest Neighbors)

O *k-nearest neighbors* (kNN) é um algoritmo não paramétrico simples de aprendizado de máquina supervisionada que pode ser usado para resolver problemas de classificação (JIANG et al., 2012). O funcionamento desse algorítimo se baseia em predizer qual a classe correta para a amostra em teste por meio do cálculo da distância entre essa amostra e as amostras de treinamento. Isso é feito por meio da escolha dos k pontos mais próximos, e calculando a probabilidade da amostra em teste pertencer a cada uma das classes das amostras de treinamento, escolhendo a de maior probabilidade como resultado da predição. A Equação 2.2 mostra seu funcionamento de forma matemática, onde S se refere à pontuação com respeito a $S(x, Cj)$, em que i é um candidato à classe j, e a saída $f(x)$ é a classe para o elemento em teste.

$$f(x) = argmax_j S(x, C_j) = \sum_{d_i \in kNN} sim(x, d_i) y(d_i, C_j) \tag{2.2}$$

É um algoritmo fácil de se implementar e se adapta a quase qualquer situação, além de conseguir resolver problemas de classificação multiclasse. As suas maiores limitações se encontram nos fatos de que à medida que a quantidade de dados em uso cresce, mais lento o algoritmo fica, e de que sua aplicabilidade depende do encontro de uma função de distância que faça sentido dentro do âmbito do problema (KOWSARI et al., 2019). Quanto maior o valor de k, maior é a confiabilidade, porém, maior é a quantidade de processamento necessária. A Figura 6 ilustra um exemplo do funcionamento desse algoritmo para um conjunto de dados bidimensional com três classes, onde os cinco vizinhos mais próximos ($k = 5$) do elemento x_i são encontrados.

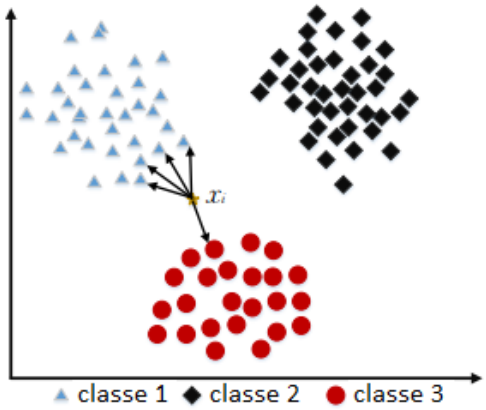

Figura 6 – Exemplificação do funcionamento do algorítimo *kNN*.

Fonte: adaptado de (KOWSARI et al., 2019).

2.4.4.2 Support Vector Machine (SVM)

O *Support vector Machine* (SVM) é um algoritmo de aprendizado de máquina supervisionada que pode ser usado para resolver problemas de classificação. O núcleo do seu funcionamento consiste na criação de hiperplanos no espaço n-dimensional do problema para separar os dados em classes, onde *n* é o número de *features* (NOBLE, 2006). Para isso, os pontos de ambas as classes mais próximos do suposto hiperplano são chamados de vetores de suporte, sendo calculada a distância entre os vetores de suporte e esse hiperplano. Essa distância é chamada de margem e o objetivo do algoritmo é maximizar a margem de forma com que a separação seja feita da melhor forma possível. Esse algoritmo pode ser usado tanto pra resolver problemas lineares quanto não lineares, como mostrado na Figura 7, sendo capaz de criar hiperplanos curvados.

Como esse algoritmo funciona por meio da separação de pontos no espaço acima e abaixo do hiperplano de classificação, não há explicação probabilística para seus resultados.

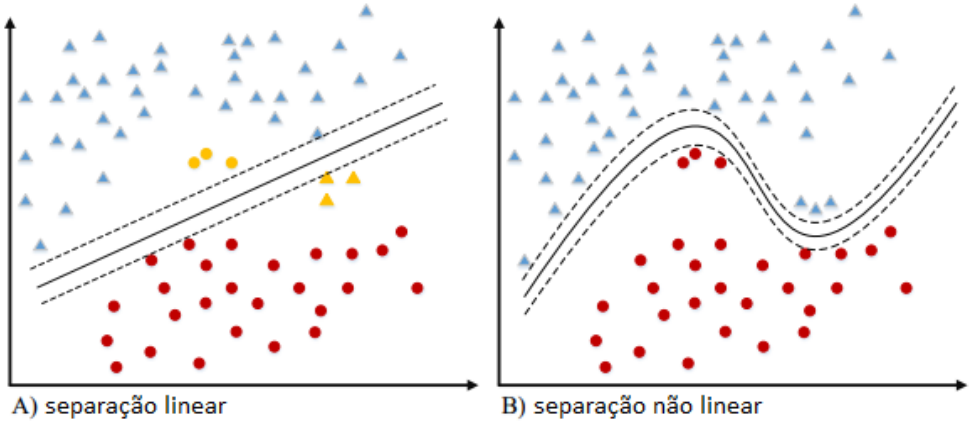

Figura 7 – Exemplificação do funcionamento do algorítimo SVM.

Fonte: adaptado de (KOWSARI et al., 2019).

2.4.4.3 *Nearest Centroid*

O *Nearest Centroid* é um algoritmo de aprendizagem de máquina que pode ser usado para solucionar problemas de classificação. O seu princípio de aplicação é extremamente simples: cada classe do problema é representada por seu centroide, que é calculado pelo valor médio de cada uma das amostras de treinamento, e cada amostra de teste é classificada de acordo com a classe do centroide mais próximo. A Figura 8 ilustra seu funcionamento (MCINTYRE; BLASHFIELD, 1980).

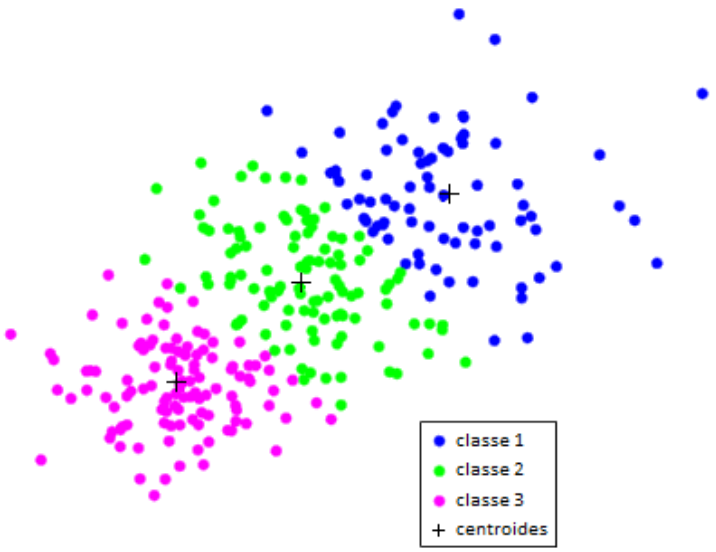

Figura 8 – Representação do funcionamento do algorítimo *Nearest Centroid*.

Figura do autor.

2.4.4.4 *Naïve Bayes Classifiers*

A família de métodos de classificação *Naïve Bayes* é um conjunto de algoritmos de aprendizado de máquina supervisionado baseados na aplicação do Teorema de Bayes com a

suposição “ingênua” de independência condicional entre cada par de *features* dado o valor da variável de classe (SARITAS; YASAR, 2019). Um classificador dessa família assume que a presença de uma característica particular em uma classe não está relacionada à presença de qualquer outra característica (KOWSARI et al., 2019). No caso em que o número de documentos n se encaixe em k categorias, onde $k \in \{c_1, c_2, ..., c_k\}$ e a classe predita como saída é $c \in C$, o algoritmo *Naïve Bayes* pode ser escrito como mostra a Equação **??**.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{2.3}$$

onde d se refere a documentos, c a classes e:

P(d) é a probabilidade anterior referente a ser o documento;
P(c) é a probabilidade anterior referente a ser a classe;
P(c|d) é a probabilidade posterior do de ser a classe visto que é o documento;
P(d|c) é a probabilidade posterior de ser o documento visto que é a classe.

contudo, a forma como esse modelo é mais utilizado é a mostrada na Equação 2.4.

$$P(c_j|d_i;\hat{\theta}) = \frac{P(c_j|\hat{\theta})P(d_i|c_j;\hat{\theta}_j)}{P(d_i|\hat{\theta})} \tag{2.4}$$

Embora os métodos derivados variem em forma, a ideia central por trás é a mesma, ou seja, assumir que o recurso satisfaça uma determinada distribuição, estimar os parâmetros da distribuição e obter a função de densidade de probabilidade.

Multinomial Naïve Bayes

Novamente, caso o número de documentos n se encaixe em k categorias onde $k \in \{c_1, c_2, ..., c_k\}$ e a classe predita como saída é $c \in C$, o algoritmo *Multinomial Naïve Bayes* pode ser escrito como na na Equação 2.5.

$$P(c|d) = \frac{P(c)\prod_{w\in d} P(d|c)^{n_{wd}}}{P(d)} \tag{2.5}$$

onde n_{wd} representa o número de vezes que a palavra w aparece no texto e $P(w|c)$ é a probabilidade de observar a palavra w dada a classe c. $P(w|c)$ é calculado como apresentado na Equação 2.6.

$$P(w|c) = \frac{1 + \sum_{d\in D_c} n_{wd}}{k + \sum_{w'} \sum_{d\in D_c} n_{w'd}} \tag{2.6}$$

Bernoulli Naïve Bayes

Assume que os dados estão distribuídos de acordo com distribuições multivariadas de Bernoulli, ou seja, pode haver vários recursos, mas cada um é assumido como um valor binário, conforme indicado na Equação 2.7.

$$P(d_i|c) = P(i|c)d_i + (1 - P(i|c))(1 - d_i) \tag{2.7}$$

Complement Naïve Bayes

É uma adaptação do algoritmo padrão*Multinomial Naïve Bayes* que é particularmente adequada para conjuntos de dados desbalanceados, como mostrado na Equação 2.8.

$$\hat{c} = argmin_c \sum_i t_i W_{ci} \tag{2.8}$$

2.4.4.5 *Passive Aggressive*

O *Passive Aggressive* é um algoritmo de aprendizagem de máquina *online* que pode ser usado para solucionar problemas de classificação. O seu princípio de funcionamento se baseia em agir de forma diferente conforme o resultado de suas predições, respondendo de forma passiva em predições corretas e agressiva em incorretas. Essa resposta passiva ou agressiva significa se será ou não produzida alguma mudança no modelo. Algoritmos com aprendizado *online* são treinados de forma sequencial, o qual o modelo é atualizado a medida que novas amostras de dados vão chegando, sendo melhor utilizados em situações que envolvam o recebimento de dados em um fluxo contínuo (CRAMMER et al., 2006).

2.4.4.6 *Random Forest*

O *Random Forest* é um algoritmo de aprendizado de máquina supervisionada que pode ser usado para resolver problemas de classificação. Seu funcionamento se baseia na criação de várias árvores de decisão aleatórias, os quais tem os seus resultados combinados para obter uma predição mais estável e precisa. O rótulo na saída do algoritmo é o escolhido pela maioria das árvores que compõe a floresta (BELGIU; DRĂGUŢ, 2016). Esse comportamento, que é ilustrado na Figura 9, é baseado no conceito de *ensembling learning*, que prega a união de vários métodos de classificação de forma a produzir resultados de melhor desempenho. A Equação 2.9 representa matematicamente seu funcionamento, onde i são as árvores de decisão, j as classes presentes nos dados e $f(x)$ é a saída do algoritmo.

$$f(x) = argmax_j \sum_{i=1}^{I} DecisionTree_{i,j} \tag{2.9}$$

A principal limitação desse algoritmo é que um elevado número de árvores torna o algoritmo muito lento e ineficaz para predições em tempo real. Em geral, esses algoritmos

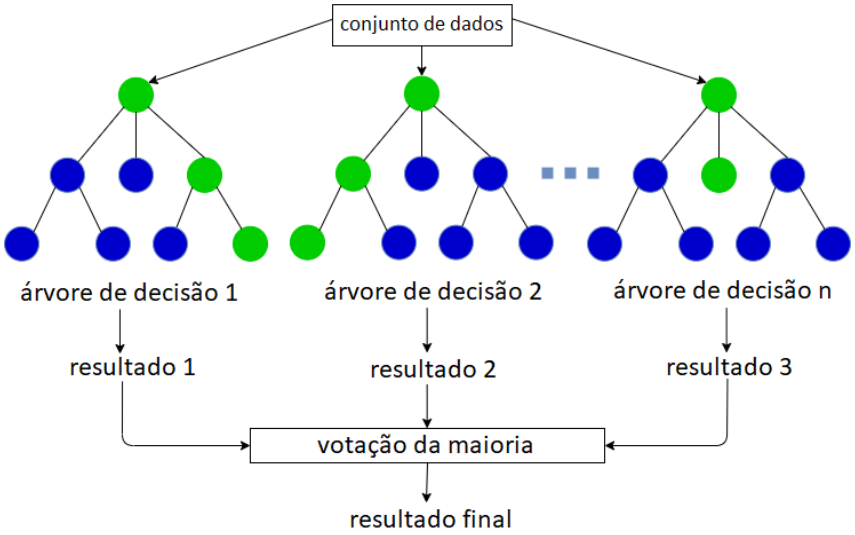

Figura 9 – Representação do funcionamento do algorítimo *Random Forest*.

Fonte: adaptado de (AMPADU, 2021).

são rápidos no treinamento e bem lentos nas predições, o que acaba virando um *trade-off*, visto que uma predição melhor requer mais árvores, o que também o deixa mais lento.

2.4.4.7 *Ridge Classifier*

O *Ridge Classifier*, que é baseado no método de regressão com mesmo nome, converte os dados do rótulo em [-1, 1] e resolve o problema com o método de regressão. O valor mais alto na predição é aceito como uma classe de destino e, para dados multiclasse, a regressão multi-saída é aplicada. A *Ridge Regression* é um tipo popular de regressão linear regularizada que inclui uma penalidade *L2*, de forma a diminuir os coeficientes para as variáveis de entrada que não contribuem muito na predição, sendo que esses coeficientes minimizam a soma residual dos quadrados (RIFKIN; LIPPERT, 2007). O problema que a regressão de Ridge busca resolver pode ser descrito matematicamente como:

$$\min_{w} ||Xw - y||_2^2 + \alpha||w||_2^2 \tag{2.10}$$

E a perda é dada pela Equação 2.11:

$$loss = sumi = 0ton(y_i \lambda yhat_i)^2 \tag{2.11}$$

Sendo que a penalidade L2 tem a forma indicada na Equação 2.12:

$$L2_{penalty} = \sum_{j=0}^{p} \beta_j^2 \tag{2.12}$$

Um hiperparâmetro λ é utilizado para ponderar a penalidade imposta sobre a função de perda. Um valor padrão de 1 impõe totalmente a penalidade, enquanto um valor de 0 desativa a penalidade, como indicado na Equação 2.13.

$$Ridge\_loss = loss + (\lambda * L2_{penalty}) \tag{2.13}$$

2.4.4.8 *Stochastic Gradient Descent* (SGD)

O *Stochastic Gradient Descent* (SGD) é um algoritmo de otimização eficiente e de fácil implementação que é utilizado em conjunto com alguns classificadores lineares (muitas vezes o SVM ou o *logistic regression*) de forma a encontrar valores de parâmetros e/ou coeficientes de funções que minimizam uma função de custo (BOTTOU, 2012). O valor mínimo para a função de custo desses classificadores não pode ser calculado diretamente, então é utilizado o SGD de forma a minimizá-lo (GARDNER, 1984). O uso desse algoritmo é muito útil em problemas de grande escala.

2.4.5 Conjunto de Dados - Treinamento, Validação e Teste

Em posse dos dados e com os métodos escolhidos, é necessário realizar a criação do conjunto de dados que será utilizado na classificação. Para isso, os dados serão divididos em três grupos com funções diferentes, os quais são descritos a seguir:

- **Treinamento:** é a amostra usada para ajustar os modelos ao conjunto de dados;
- **Validação:** é a amostra de dados usada para fornecer uma estimativa imparcial da capacidade de predição dos modelos em cima da amostra de treinamento enquanto são ajustados seus hiperparâmetros;
- **Teste:** é a amostra usada para avaliar o ajuste final dos modelos ao conjunto de dados.

A proporção ótima para essa divisão depende do problema a ser resolvido e pode ser escolhida através do uso de um procedimento computacional que usa regressão logística, sendo usualmente de 70-80% para treinamento, e 20-30% para validação e teste (AFENDRAS; MARKATOU, 2019). De forma a não diminuir a quantidade de dados disponível para ser usada em treinamento e teste, a validação pode ser feita por meio do uso da técnica de validação cruzada, que é descrita na próxima subseção.

2.4.5.1 Validação Cruzada (*k-Fold Cross-Validation*)

A validação cruzada, em inglês *k-fold cross-validation*, é uma técnica utilizada principalmente em aprendizado de máquina para estimar o ajuste do modelo em cima da amostra de treinamento em problemas com dados escassos ou para avaliar um fenômeno de baixa

ocorrência (BROWNE, 2000). A técnica consiste em particionar os dados *k* vezes de forma a reutilizar a mesma amostra no treinamento dos modelos. Essa técnica previne o *overfitting* e tende a diminuir o erro de predição dos modelos (BROWNE, 2000).

Cada elemento da amostra de dados é atribuído a uma partição e lá permanece durante todo o procedimento, o que faz com que cada um deles seja utilizado uma vez para testar o modelo e *k-1* vezes para treinar. A escolha do *k* ótimo pode também ser feita por meio do uso de um procedimento computacional que usa regressão logística (AFENDRAS; MARKATOU, 2019) A Figura 10 mostra um esquema de como é utilizada a técnica *10-fold cross-validation* nos processos de treinamento e teste dos dados, que é exatamente o que será feito neste trabalho.

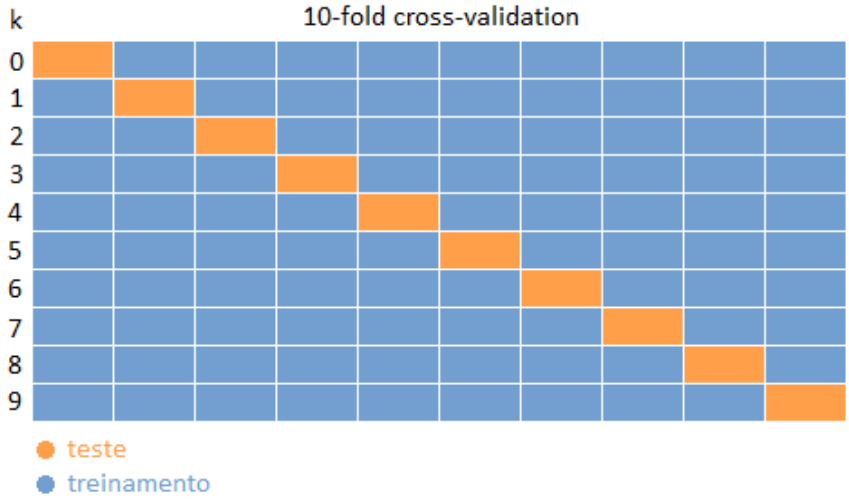

Figura 10 – Esquemático representando o uso da técnica *10-fold cross-validation*.

Figura do autor.

É importante dizer que para fazer o cálculo de qualquer métrica de desempenho de forma a avaliar os processos de classificação multiclasse que utilizam essa técnica é necessário fazer uma média entre os resultados de todas as classes. O funcionamento dessa técnica pode ser descrito pelo fluxo de trabalho (itens) abaixo:

1. Misture o conjunto de dados;
2. Escolha o *k* desejado;
3. Divida o conjunto de dados em *k* partições de forma aleatória;
4. Para cada partição:
  4.1. Separe essa partição como amostra para teste;
  4.2. Agrupe as *k-1* partições restantes como amostra para treinamento;
  4.3. Realize o procedimento de classificação;
  4.4. Calcule as métricas de desempenho para a partição;
5. Calcule as métricas de desempenho gerais.

2.4.5.2 Balanceamento de Classes

A diferença entre o número de amostras de cada uma das classes é o que define se um conjunto de dados está balanceado ou não. Para se ter classes balanceadas é preciso ajustar o número de amostras de todas as classes para o número referente à classe com menos amostras. Isso garante que os modelos utilizados consigam generalizar bem sua predição de forma com que os resultados da classificação não sejam polarizados para o lado das classes com maior número de amostras (POOLSAWAD; KAMBHAMPATI; CLELAND, 2014).

É importante dizer que existem casos nos quais seja importante ter classes desbalanceadas. Nesses casos a frequência de ocorrência de algumas classes é muito maior no problema real, e se a quantidade de amostras for limitada de forma a prejudicar os resultados da classificação caso o balanceamento seja feito, é melhor realizar o processo com classes desbalanceadas de tal maneira que o modelo tenha acesso a uma maior quantidade de amostras referentes às classes mais frequentes e produza melhores resultados. A escolha entre como deve ser feito o balanceamento e principalmente se ele deve ser feito em primeiro lugar depende apenas do problema em questão.

2.4.6 Avaliação dos Resultados

A melhor forma de avaliar os resultados produzidos por modelos de classificação é por meio do uso de indicadores de desempenho, e isso é normalmente feito através da utilização de métricas estatísticas. Várias delas estão associadas aos conceitos por trás da matriz de confusão, sendo necessário entender o que ela é e como se relaciona ao cálculo dessas métricas. A teoria por trás da matriz de confusão é apresentada na próxima subseção e em seguida é descrito tudo o que será utilizado na avaliação dos resultados.

2.4.6.1 Matriz de Confusão

A matriz de confusão é uma tabela que indica o número de ocorrências de um valor que se enquadra nas condições impostas por dois rotuladores: o valor real e o predito pelo modelo de classificação em questão (GRANDINI; BAGLI; VISANI, 2020). A Figura 11 mostra a estrutura de uma matriz de confusão associada a um problema binário, onde as colunas representam as predições e as linhas os valores reais. A diagonal principal (em verde) corresponde às ocasiões em que os dois rotuladores concordam entre si, enquanto a diagonal secundária (em vermelho) às ocasiões no qual existem erros de predição. Os possíveis tipos de ocorrências para a matriz de confusão em questão são:

- **Verdadeiros Positivos (VP):** quando o modelo prevê o resultado como positivo e ele é de fato positivo, consistindo de uma previsão correta;
- **Verdadeiros Negativos (VN):** quando o modelo prevê o resultado como negativo e ele é de fato negativo, consistindo de uma previsão correta;

- **Falsos Positivos (FP):** quando o modelo prevê o resultado como positivo e ao invés disso ele é negativo, consistindo de um erro na previsão;
- **Falsos Negativos (FN):** quando o modelo prevê o resultado como negativo e ao invés disso ele é positivo, consistindo de um erro na previsão.

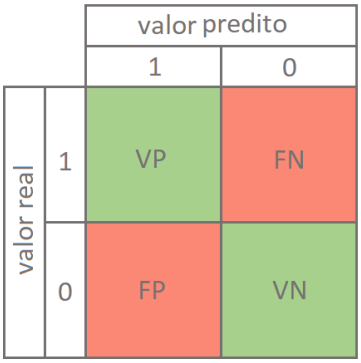

Figura 11 – Estrutura de uma matriz de confusão para classificação binária.

Figura do autor.

Em problemas de classificação multiclasse as matrizes de confusão são um pouco mais complexas. As linhas e colunas não são mais valores e sim classes, e as ocorrências passam a se referir a uma classe, não estando mais restritas a somente uma posição da tabela. É como se existissem várias matrizes de confusão binárias dentro de uma matriz maior. A Figura 12 mostra a estrutura de uma matriz de confusão associada a um problema multiclasse.

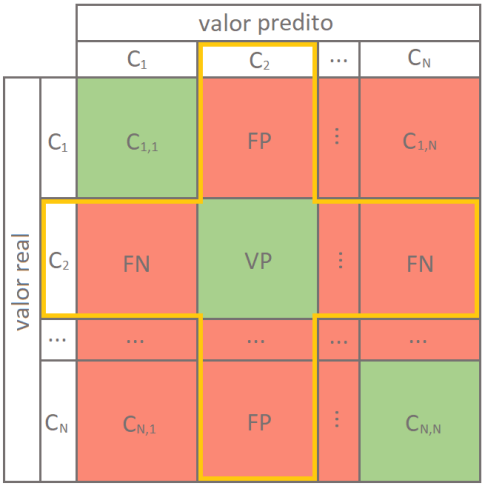

Figura 12 – Estrutura de uma matriz de confusão para classificação multiclasse.

Figura do autor.

A quantidade de verdadeiros positivos de uma classe é dada pelo valor da célula que se encontra no cruzamento entre a linha e a coluna referentes à classe em questão. A quantidade de falsos positivos de uma classe é dada pela soma de todos os valores das outras

células presentes na linha referente à classe em questão. A quantidade de falsos negativos de uma classe é dada pela soma de todos os valores das outras células presentes na coluna referente à classe em questão, e a quantidade de falsos negativos é dada pela soma dos valores de todas as demais células restantes.

O problema de classificação de cada classe pode então ser tratado como um problema binário e o cálculo de métricas estatísticas individuais para cada uma das classes pode ser feito. Para se ter uma visualização geral do desempenho de um modelo é necessário o cálculo de médias entre as métricas de todas as classes, produzindo um valor só que represente o modelo como um todo. Essas médias podem ser feitas de diferentes formas: o *macro-average* é calculado pela média aritmética simples das métricas de todas as classes, tratando todas as classes de forma igual sem se importar com o balanceamento, já o *micro-average* é calculado a partir da soma da quantidade de ocorrências de todas as classes juntas, tratando cada predição individual de forma igualitária.

2.4.6.2 Métricas Estatísticas

A escolha das métricas adequadas à avaliação dos resultados de um modelo depende muito do problema a ser resolvido, sempre existindo alguma que se sobressai em situações específicas e não vai tão bem em outras (GRANDINI; BAGLI; VISANI, 2020). As métricas a serem utilizadas na avaliação dos resultados deste trabalho são derivadas da matriz de confusão, sendo todas elas descritas a seguir. Para essa descrição, faz se necessário a especificação das seguintes variáveis:

- **VP:** verdadeiros positivos
- **VN:** verdadeiros negativos
- **FP:** falsos positivos
- **FN:** falsos negativos
- **A:** acurácia
- **P:** precisão
- **S:** sensibilidade
- **E:** especificidade
- **F1:** F1-*score*
- **T:** número total de amostras

Acurácia

A acurácia é uma métrica que avalia o modelo quanto a sua capacidade de acerto, sendo utilizada como uma medida de desempenho geral. Essa métrica precisa ser usada com cuidado, visto que nem sempre é a melhor representação do quão bom um modelo é na

resolução de um problema (GRANDINI; BAGLI; VISANI, 2020). A Equação 2.14 mostra como essa métrica é calculada:

$$A = \frac{VP + VN}{T} \tag{2.14}$$

Precisão

A precisão é uma métrica que avalia o modelo quanto a sua capacidade de acerto nos casos rotulados como positivos. Essa métrica é utilizada em casos nos quais falsos positivos são considerados mais prejudiciais à situação em questão do que falsos negativos (GRANDINI; BAGLI; VISANI, 2020). A Equação 2.15 mostra como essa métrica é calculada:

$$P = \frac{VP}{VP + FP} \tag{2.15}$$

Sensibilidade e Especificidade

A sensibilidade, ou *recall*, é uma métrica que avalia o modelo quanto a capacidade de detecção de resultados classificados como positivos. Essa métrica é utilizada em casos onde falsos negativos são considerados mais prejudiciais à situação em questão do que falsos positivos (GRANDINI; BAGLI; VISANI, 2020). A Equação 2.16 mostra como essa métrica é calculada:

$$S = \frac{VP}{VP + FN} \tag{2.16}$$

A especificidade é o oposto da sensibilidade, sendo uma métrica que avalia o modelo quanto a capacidade de detecção de resultados classificados como negativos (GRANDINI; BAGLI; VISANI, 2020). A Equação 2.17 mostra como essa métrica é calculada:

$$E = \frac{VN}{VN + FP} \tag{2.17}$$

F1-*Score*

O F1-*score* é uma métrica bem específica calculada através da média harmônica entre precisão e sensibilidade. Ela é muito útil em casos quando ambas as métricas são importantes na análise de determinada situação, possibilitando a representação de um conjunto de dados por um único valor (GRANDINI; BAGLI; VISANI, 2020). A Equação 2.18 mostra como essa métrica é calculada:

$$F1 = \frac{2 \times P \times S}{P + S} \tag{2.18}$$

2.4.6.3 Curva ROC

A curva ROC, que é acrônimo do inglês: *Receiver Operating Characteristic*, ou curva característica de operação do receptor na tradução direta, é uma métrica gráfica de desempenho utilizada para avaliar a qualidade dos resultados de um processo de classificação (BRADLEY, 1997). Ela é representada em dois eixos: no eixo Y está a taxa de verdadeiros positivos (TVP), ou sensibilidade, que é dada pelo número de verdadeiros positivos dividido pelo total de amostras esperadas como verdadeiras; e no eixo X está a taxa de falsos positivos (TFP), ou especifidade, que é dada pelo número de falsos positivos dividido pelo total de amostras esperadas como negativas. Com isso, o gráfico produz em seu canto superior esquerdo um ponto considerado ideal, que representa uma TFP igual a zero e uma TVP igual a um. A AUC, que é uma sigla do inglês: *Area Under the Curve*, ou área abaixo da curva na tradução direta, pode então ser calculada para avaliar a qualidade do classificador, com áreas maiores correspondendo a resultados melhores. A inclinação da curva também tem um significado importante, considerando que é sempre ideal maximizar a TVP e minimizar a TFP (BRADLEY, 1997).

Essas curvas são, geralmente, utilizadas para avaliar problemas de classificação binária, mas podem também ser empregadas em classificação multiclasse, sendo necessário ou representar cada classe por sua própria curva ou considerar cada predição de classe como binária por meio do cálculo da média entre os valores de todas as classes (*micro/macro-average*), como já mencionado anteriormente. O gráfico exibido na Figura 13 exemplifica essa situação por meio da mostra de curvas ROC derivadas de um problema de classificação multiclasse.

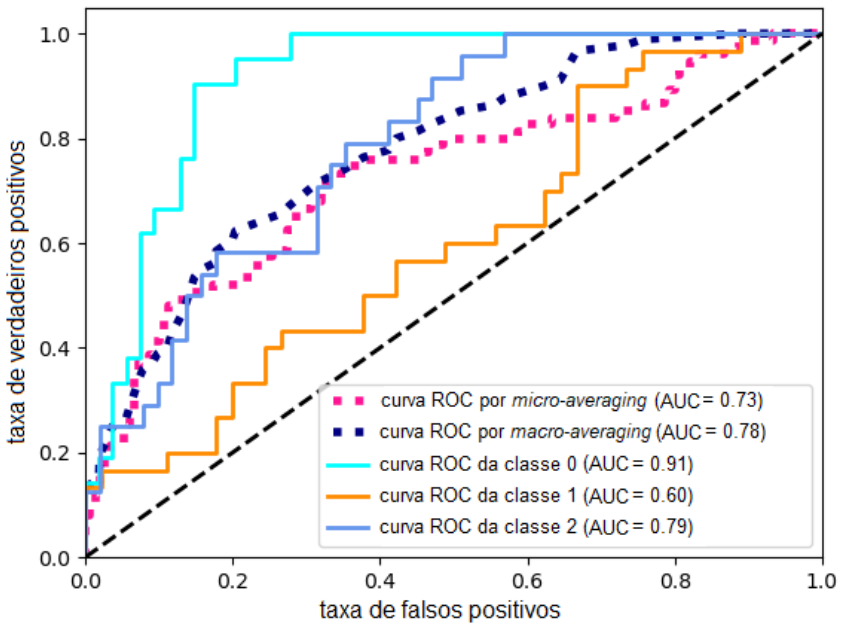

Figura 13 – Exemplos de curvas ROC derivadas de um problema de classificação multiclasse.

Fonte: adaptado de (SCIKIT-LEARN, s.d.).

3 Trabalhos Relacionados

Neste capítulo é descrito o *Deep Vacuity*, que é um projeto maior no qual este trabalho está inserido e que serve de motivação e inspiração para o mesmo. São também mencionados outros trabalhos relevantes que tem alguma conexão com o que será feito aqui, priorizando os que envolvem processamento de linguagem natural e dando ênfase para os que fazem o uso de textos públicos e/ou realizam classificação de texto.

3.1 Projeto Deep Vacuity

O *Deep Vacuity* (DV), que foi devidamente apresentado em (LIMA, 2021), é um projeto resultante de um convênio entre o Instituto Nacional de Criminalística (INC), que é o órgão central de criminalística da PF, e da Universidade de Brasília (UnB). Esse projeto tem como principal objetivo o combate à corrupção por meio do desenvolvimento de metodologias para identificação de formação de cartel em obras públicas utilizando técnicas de aprendizado de máquina e inteligência artificial. Com isso, o projeto busca auxiliar em tarefas de fiscalização, auditoria e investigação, as quais atualmente são em sua grande maioria feitas de forma manual por análise humana. Assim, de forma a dar suporte a todas essas atividades está sendo desenvolvido um sistema computacional[1] com interface de usuário que pretende ter as seguintes funcionalidades:

- Visualização de dados, metadados, agrupamentos e seleção de entidades;
- Disponibilização de um painel de controle com gráficos e informações em escala macro e micro de determinados campos do conhecimento;
- Visualização de grafos e correlações entre dados, documentos e entidades;
- Representação espacial bidimensional e/ou tridimensional das relações entre dados;
- Visualização de dados em diferentes escalas e tipos de intervalos temporais;
- Representação de dados por geolocalização;
- Capacidade de interação com o usuário para validação e/ou refinamento de resultados de modelos funcionais inteligentes;
- Capacidade de retroalimentação de dados para retreinamento de modelos funcionais inteligentes já existentes ou realizar treinamentos completamente novos.

[1] Esse sistema pode ser acessado pelo endereço: `https://deepvacuity.cic.unb.br/`

O projeto está em andamento e possui uma equipe qualificada formada por profissionais de ambos os conveniados que estão trabalhando de forma interdependente e coordenada a fim de não só atingir os objetivos mas também se ater ao planejamento inicial.

O sistema computacional que está em construção fará a gestão do banco de dados e cuidará da aplicação das metodologias e técnicas desenvolvidas e dos recursos relacionados à visualização de dados. Além disso, o sistema deve possibilitar que tudo isso seja acessado por meio de uma interface com o usuário. A Figura 14 mostra a arquitetura inicial proposta para o sistema, na qual é apresentado um referencial para sua futura infraestrutura e organização, e que busca atender todas as expectativas para suas funcionalidades. Essa ilustração foi elaborada por Leonardo Carvalho como parte do Relatório Descritivo de Atividades e Produtos Desenvolvidos no projeto (esse documento não foi publicado).

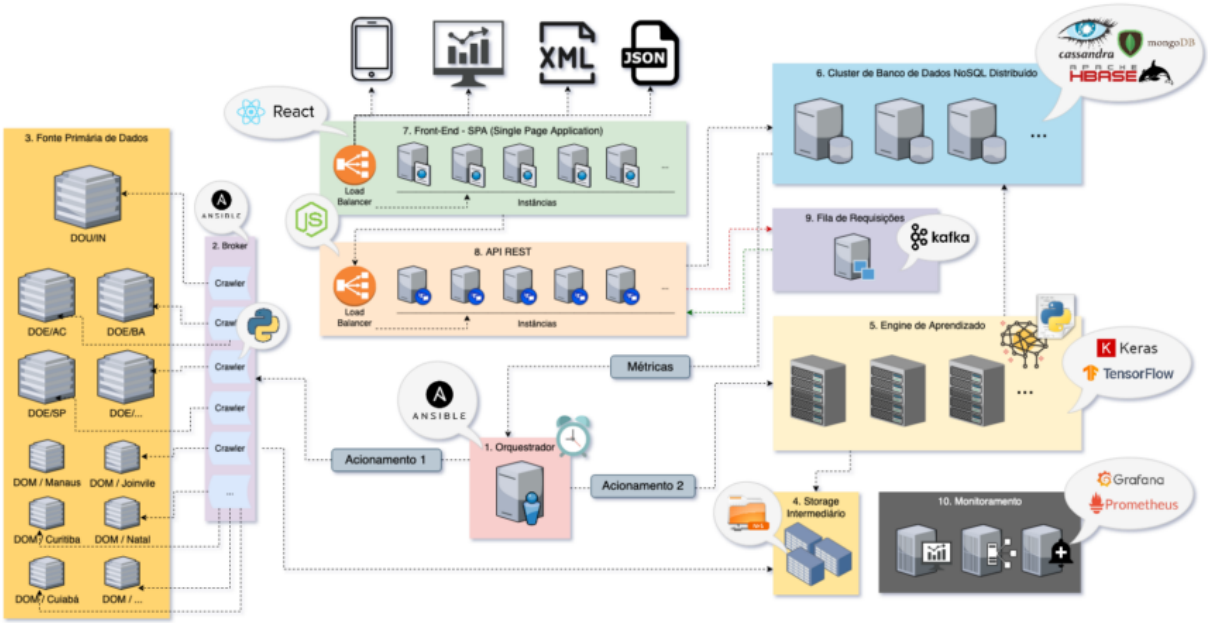

Figura 14 – Arquitetura inicial proposta para o sistema do projeto *Deep Vacuity*.

Fonte: Lima (2021, p. 42).

Um dos primeiros trabalhos que fez parte do DV foi desenvolvido por Marcos Cavalcanti Lima em (LIMA, 2021), no qual foi definido um modelo de classificação de publicações do Diário Oficial da União como forma de detectar indícios de fraudes e conluios em licitações de obras públicas no Brasil. Os resultados desse trabalho foram bem expressivos, visto que foi o pioneiro na resolução desse tipo de problema e obteve excelentes resultados.

Este trabalho fará parte do DV como uma das várias metodologias de combate à corrupção que estarão inclusas no sistema computacional final. Assim, o método de classificação descrito na Seção 1.2, será adaptado de forma a fazer parte do sistema como uma ferramenta disponível para uso quando necessário.

3.2 Outros Trabalhos Relevantes

O primeiro trabalho a ser mencionado é a pesquisa sobre algoritmos de classificação de texto feita em (KOWSARI et al., 2019), de 2019. Nesse trabalho foi feito um breve apanhado geral sobre o tema, criando uma discussão acerca de alguns algoritmos de classificação de texto existentes: *Rocchio Algorithm*, *Bagging and Boosting*, *Logistic Regression* (LR), *Naïve Bayes Classifier* (NBC), *k-Nearest Neighbor* (kNN), *Support Vector Machine* (SVM), *Decision Tree Classifier* (DTC), *Random Forest*, *Conditional Random Field* (CRF) e alguns outros que utilizam aprendizagem profunda. O trabalho também cobre métodos de extração de características, redução de dimensionalidade e de avaliação de resultados, sendo ainda feitas algumas discussões sobre as limitações desses algorítimos e sobre suas aplicações no mundo real. O trabalho é excepcional e acaba sendo uma leitura obrigatória para interessados.

Outro trabalho de 2019 foi o feito de classificação multiclasse de texto de larga escala em documentos legais europeus, criando um novo conjunto de dados constituído de 57 mil documentos legislativos anotados, e fazendo testes com diversos classificadores neurais. Os melhores resultados foram obtidos com o uso do algoritmo BIGRUs, tendo desempenho melhor do que outros algoritmos do estado da arte (CHALKIDIS et al., 2019).

Também em 2019 foi feito um trabalho comparativo do desempenho de classificação entre os algoritmos *Multinomial Naïve Bayes* e *Bernoulli Naïve Bayes*, que são os dois mais populares da família, em um problema que visava predizer se o sentimento de um artigo de notícias era positivo ou negativo (SINGH et al., 2019). O trabalho obteve resultados medianos devido à baixa quantidade de dados, mas observou que o *Multinomial* teve desempenho levemente melhor na tarefa.

Em 2020, no Brasil, foi desenvolvido um modelo baseado em casos CBR, sigla do inglês *Case-Based Reasoning*) abordando os processos de licitações para pavimentação de áreas urbanas. Estes serviços demandam uma grande quantidade de recursos públicos, dando muita margem e sendo constantemente alvo de ações criminais. O modelo se baseou em uma análise estruturada de dados que considerava: o tipo de licitação, as empresas envolvidas, os contratos e os dados de localização. Assim sendo, foram classificados então casos de conluio (VALLIM, 2020). Calculou-se então um indicador de conluio a partir das projeções e confrontos realizados, sendo os resultados obtidos bem satisfatórios. Isso levou à determinação de que a metodologia desenvolvida pode ser replicada e utilizada para fins de controle de contas em contratos de obras públicas, para ações policiais de repressão a crimes contra à economia e de fraudes às licitações.

Também em 2020, foi apresentado o VICTOR (DE ARAUJO et al., 2020), que é um novo conjunto de dados feito para classificação de documentos legais brasileiros, sendo desenvolvido também por pesquisadores da Universidade de Brasília. O mesmo conta com 692 mil documentos e suporta dois tipos de tarefas: classificação quanto ao tipo de documento

e classificação por temática do documento.

Em 2021, Aguiar e Silveira (AGUIAR et al., 2021) investigaram diferentes métodos de classificação de texto e diferentes combinações de *word embeddings* na tarefa de classificar 16 mil petições iniciais e indiciamentos do Tribunal de Justiça do Estado do Ceará, no Brasil. Esses processos foram classificados nas cinco classes mais representativas do Processo Civil Comum: Execução de Título Extrajudicial, Ação Criminal - Processo Ordinário, Processo Civil Especial e Execução Fiscal. O trabalho obteve resultados razoáveis alcançando uma nota de 0.88 para o macro F1-*score* para o melhor caso utilizando o modelo.

4 Metodologia Proposta

Neste capítulo o caminho percorrido ao longo da realização do trabalho é descrito de forma detalhada, com cada seção correspondendo diretamente a uma etapa do desenvolvimento. Assim, o diagrama de fluxo mostrado na Figura 15 representa essas etapas, que incorporam o *pipeline* de classificação de texto apresentado na Seção 2.4.

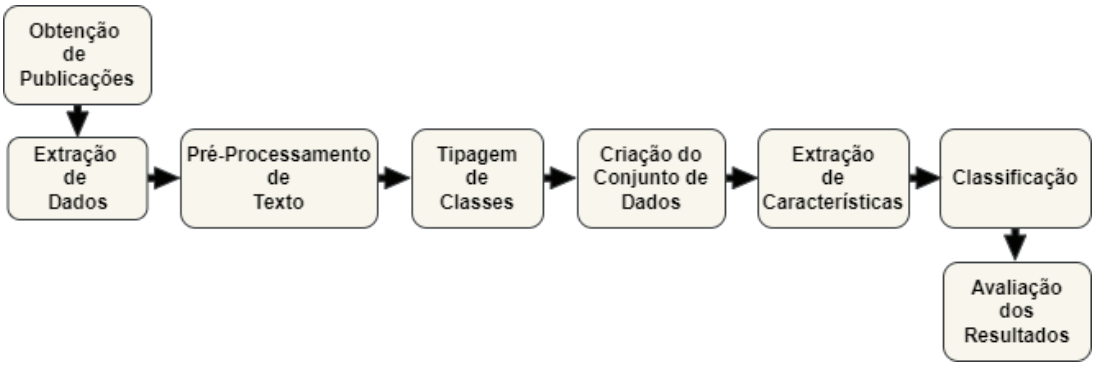

Figura 15 – Diagrama de fluxo representando as etapas de desenvolvimento do trabalho.

Figura produzida pelo autor.

4.1 Obtenção de Publicações

Os dados necessários ao projeto, que precisam estar em grande quantidade, consistem em textos de publicações, identificadores que façam com que cada publicação seja única em relação às demais, e rótulos que indiquem o tipo de ato o qual cada publicação representa. Publicações sem rótulo ou identificador fariam com que a carga de trabalho fosse bem maior, visto que esses processos teriam então que ser feitos manualmente, já publicações sem texto são descartáveis. Lembrando que essas publicações têm origem na Seção 3 do DOU.

A fim de reunir um número grande o suficiente de dados, a obtenção de publicações será feita de duas formas. A primeira a partir de um acervo da própria PF, no qual há vários arquivos textuais em formato JSON com centenas de publicações cada. A segunda será feita por meio do servidor do projeto DV, que possui anos de arquivos textuais de publicações retiradas diretamente do DOU, armazenadas também em formato JSON. Todas as publicações a serem obtidas pelo primeiro método possuirão identificador, texto e rótulo, já as a serem obtidas pelo segundo método podem não ter todas essas informações.

O acesso aos arquivos do acervo da PF será simples e direto, não sendo necessário nenhum procedimento especial. Já na obtenção pelo segundo método, as publicações estarão armazenadas no servidor em páginas de dados, e a obtenção será feita por meio de uma sequência de requisições, obtendo um número determinado de publicações por página até

atingir uma quantidade razoável. As páginas estão em ordem cronológica, e quanto mais antiga for a publicação, menor será a chance da mesma possuir identificador e/ou rótulo, sendo necessário dar preferência às páginas mais novas. Ao mesmo tempo, é necessário que publicações mais antigas também estejam presentes, de forma com que o conjunto de dados represente bem uma publicação de qualquer época.

4.2 Extração de Dados

Os arquivos textuais de publicações a serem obtidos possuem informação dispensável, o que cria a necessidade de serem separados somente os dados necessários ao trabalho. Assim sendo, Serão extraídos os números de identificação das publicações, seus textos e seus rótulos, sendo essas informações organizadas e estruturadas. Parte desses arquivos terão essas informações formatadas em XML, e como os mesmos estarão em formato JSON, as informações desejadas irão ser facilmente encontradas por meio de buscas utilizando palavras-chave (campos *_id*, *text* e *artType*).

Foi disponibilizado um exemplo mostrando a formatação desses arquivos no Anexo – A. Além disso, serão também removidos os dados extraídos de publicações repetidas e de publicações que não tenham todas as informações necessárias ao projeto, sendo o primeiro feito por meio da comparação dos identificadores de diferentes publicações e o segundo pela simples busca de campos vazios nos dados já extraídos e organizados.

4.3 Pré-Processamento de Texto

A presença de ruído e informações irrelevante pode influenciar negativamente no desempenho dos modelos no processo de classificação, levando à necessidade de realizar procedimentos de filtragem e transformação nos textos de publicações de forma a produzir melhores resultados. O texto passará por processos de minusculização e tokenização, além de serem removidos símbolos, pontuação, abreviações, *stop words*, marcações HTML e possíveis resquícios de alguma estrutura de dados.

O pré-processamento de texto é uma etapa padrão de trabalhos que envolvem NLP, e todos os procedimentos que o constituem foram listados e descritos na Seção 2.4.1. Considerando que essa etapa consiste em remover partes do texto, serão feitas análises para decidir quais os procedimentos entre os mencionados serão de fato úteis ao projeto, visto que a perda de alguma informação específica pode causar um impacto negativo nos resultados da classificação. Essas análises serão feitas com base na comparação entre os resultados da classificação utilizando diferentes abordagens como métodos de pré-processamento.

4.4 Tipagem de Classes

A tipagem de classes consistirá na categorização dos textos de publicações de acordo com a sua função dentro do processo licitatório, os agrupando em tipos de publicações por meio de seus rótulos e transformando esses tipos em classes. Para isso, serão priorizados os tipos com grande número de dados e os tipos considerados importantes para o trabalho de fiscalização, sendo os atos produzidos pelas etapas de um processo licitatório de extrema importância nesse caso (convênios, licitações, contratos, termos aditivos, etc). Mais informações a respeito dos tipos de atos que constituem as publicações feitas no DOU e sobre as etapas de um processo licitatório podem ser encontradas nas Seções 2.2.1 e 2.1.1, respectivamente.

Para isso, os dados organizados e estruturados serão agrupados com base em seus rótulos, o que irá permitir a criação de uma lista com todos os rótulos presentes. Além disso, será também possível realizar a contagem do número de publicações associadas a cada rótulo. Essas informações serão então utilizadas em análises quantitativas e qualitativas, cujos resultados irão permitir que rótulos semelhantes sejam convertidos em tipos de publicações. Essas análises serão feitas de forma heurística, sendo os tipos de publicações com maior relevância transformados em classes, priorizando os tipos ligados aos fatores mencionados anteriormente. Quanto mais classes, maior é o impacto positivo potencial do projeto, e quanto maior o número de publicações de cada classe, melhor e mais confiável será o resultado. A escolha das classes terá grande impacto no resultado final, fazendo com que essa etapa seja de extrema importância. Por isso, esse processo será repetido até que um equilíbrio entre o número de classes e a quantidade de publicações por classe seja encontrado, produzindo os melhores resultados possíveis, visto que o número de publicações de cada classe a ser utilizado na próxima etapa deve ser igual.

4.5 Criação do Conjunto de Dados

Os dados precisam ser reorganizados de forma a permitir que o treinamento e o teste dos modelos sejam feitos de forma adequada. A fim de que o resultado do processo de classificação não seja polarizado, os dados serão reestruturados, separados por classe e balanceados, visto que para isso o número de publicações de cada classe não pode divergir muito. Portanto, todas as classes serão limitadas a ter o mesmo número de publicações da classe com a menor quantidade. Mais informações sobre a importância e sobre os requisitos do balanceamento de classes estão na Seção 2.4.5.2.

Com as classes balanceadas, os dados de cada classe serão então divididos em 10 lotes diferentes com o mesmo número de publicações, os quais serão agrupados com os de outras classes de forma a criar 10 partições com a mesma proporção de publicações para cada classe. Isso possibilitará o treinamento e o teste dos modelos de classificação por meio do uso da técnica de validação cruzada, que é descrita na Seção 2.4.5.1.

4.6 Extração de Características

Os modelos de classificação a serem utilizados não processam informação textual, fazendo com que seja necessário representar os dados na forma numérica. Para isso, os textos de publicações serão primeiro transformados em vetores contínuos de texto e, então, passados à representação numérica utilizando a técnica TF-IDF. Esse procedimento não só fará com que os dados sejam representados numericamente como auxiliará os modelos de classificação na extração de características. Mais informações a respeito de como funciona a técnica e como serão representados os dados podem ser encontradas na Seção 2.4.2.

4.7 Classificação

Essa etapa fará a simulação do processo de classificação que seria feito por um ser humano capacitado. Para isso, foi escolhido um conjunto de 14 modelos para serem testados, sendo alguns deles pequenas modificações do outro. Esses modelos são listados a seguir:

- *k-Nearest Neighbors* (kNN);
- *Support Vector Machine* (SVM) com penalidade L1;
- *Support Vector Machine* (SVM) com penalidade L2;
- *Support Vector Machine* (SVM) com extração de características baseada em L1 e penalidade L2;
- *Nearest Centroid*;
- *Multinomial Naïve Bayes*;
- *Bernoulli Naïve Bayes*;
- *Complement Naïve Bayes*;
- *Passive Aggressive*;
- *Random Forest*;
- *Ridge Classifier*;
- *Stochastic Gradient Descent* (SGD) com penalidade *Elastic Net*;
- *Stochastic Gradient Descent* (SGD) com penalidade L1;
- *Stochastic Gradient Descent* (SGD) com penalidade L2.

Esse conjunto de modelos consiste em classificadores lineares clássicos com funcionamento baseado no uso de aprendizagem de máquina supervisionada e foram escolhidos por terem um histórico de bom desempenho na resolução de problemas semelhantes ao deste trabalho. Essa escolha foi feita com base no software apresentado por (PEDREGOSA et al., 2011) e as características e informações acerca do funcionamento de cada um deles podem ser encontradas na Seção 2.4. As configurações e parâmetros dos modelos serão optimizadas por meio da comparação entre os resultados utilizando diferentes combinações.

4.8 Avaliação dos Resultados

Para avaliar os resultados serão utilizadas métricas calculadas a partir da matriz de confusão multiclasse, conforme foi descrito na Seção 2.4.6. Essas métricas serão retiradas do processo de classificação de cada um dos modelos e serão comparadas entre si para a escolha do melhor desempenho. As métricas de desempenho estatístico a serem utilizadas são: acurácia, precisão, sensibilidade e F1-*score*. Os valores finais para cada métrica serão os encontrados no melhor caso de validação cruzada, sendo os mesmos apresentados em tabelas e gráficos de forma a facilitar a visualização e análise dos resultados. Considerando o que foi exposto na Seção **??**, os valores de F1-*score* terão um peso maior na avaliação.

5 Resultados e Discussões

Neste capítulo os resultados do trabalho são apresentados, discutidos e analisados, sendo também levantadas algumas das dificuldades encontradas em seu desenvolvimento e explicados alguns aspectos do processo de tomada de decisões. De forma a organizar melhor o que será exposto, este capítulo foi subdividido em três seções. A primeira se restringirá a apresentar os resultados relacionados ao conjunto de dados, a segunda os relacionados à classificação e a terceira fará análises a respeito dos resultados como um todo.

5.1 Resultados Relacionados ao Conjunto de Dados

Somando-se os números resultantes dos dois métodos de obtenção de publicações utilizados, um total de 4.941.827 publicações foram obtidas inicialmente. Após a remoção de exemplares que não poderiam ser utilizados, foram extraídos identificadores (*pub_id*), textos (*pub_text*) e rótulos (*pub_type*) dos restantes, resultando em uma amostra com 4.181.390 publicações, como é mostrado na Figura 16.

	pub_id	pub_text	pub_type
0	1347533	O DESEMBARGADOR DO TRABALHO, PRESIDENTE DO TR...	Ato
1	1368786	A Diretora-Geral do Instituto Nacional de Cân...	Portaria
2	1684288	Credenciada: Hidrofísio Centro de Reabilitaçã...	Extrato de Credenciamento
3	1708516	Credenciada: Centro de Estudo e Pesquisa Of...	Extrato de Credenciamento
4	1684291	Credenciada: Serviço Social da Indústria - SES...	Extrato de Credenciamento
...	...	...	...
4181385	13975601	AVISO DE LICITAÇÃO PREGÃO Nº 2/2018 - UASG 250...	Aviso de Licitação-Pregão
4181386	13976262	EXTRATO De AUTORIZAÇÃO Nº 13/2018 - ANATEL P...	Termo de Autorização
4181387	13976263	EXTRATO DE ACORDO DE COOPERAÇÃO Convênio 01/...	Extrato de Convênio
4181388	13976261	AVISO DE ADJUDICAÇÃO E HOMOLOGAÇÃO PREGÃO ELET...	Aviso de Homologação e Adjudicação
4181389	13976256	EXTRATOs DE CONTRATOs Objeto: Aquisição de I...	Extrato de Contrato

4181390 linhas x 3 colunas

Figura 16 – Resultados da obtenção de publicações e extração de dados.

Figura produzida pelo autor.

Na etapa de pré-processamento de texto, os procedimentos adotados foram: remoção de símbolos, pontuação, números, *stop words* e ruído, minusculização, tokenização, e lematização. Esses procedimentos em conjunto resultaram nos melhores resultados para os dados desse trabalho. Essa análise foi feita a partir da comparação entre os resultados utilizando diferentes combinações de procedimentos. Outros procedimentos foram testados, mas não resultaram em mudanças expressivas nos resultados ou foram considerados como prejudiciais, removendo informação relevante.

Os dados obtidos possuíam 243 rótulos diferentes, sendo feita uma análise de forma a agrupá-los em classes de acordo com o tipo de publicação ao qual se referiam, e com o número de publicações pertencentes a cada um. A partir dessa análise, o conjunto de dados foi então organizado para três cenários diferentes, sendo variada a quantidade de classes e de publicações por classe. Os cenários resultantes da tipagem de classes são descritos a seguir:

- Conjunto de dados com **cinco** classes e **397.000** publicações, onde cada classe possui **79.400** exemplares. As classes foram, em ordem: não definido (00), contrato (01), licitação (02), convênio (03) e termo aditivo (04);
- Conjunto de dados com **doze** classes e **95.880** publicações; onde cada classe possui **7.990** exemplares. As classes foram, em ordem: não definido (00), contrato (01), licitação (02), convênio (03), termo aditivo (04), ato administrativo (05), portaria (06), ata (07), alvará (08), intimação (09), doação (10) e concurso público (11).
- Conjunto de dados com **vinte** classes e **16.600** publicações, onde cada classe possui **830** exemplares. As classes foram, em ordem: não definido (00), contrato (01), licitação (02), convênio (03), termo aditivo (04), ato administrativo (05), portaria (06), ata (07), alvará (08), intimação (09), doação (10), cooperação (11), concurso público (12), cessão de uso (13), balanço financeiro (14), acórdão (15), consulta pública (16), parecer técnico (17), leilão (18) e circular (19).

No primeiro cenário a escolha das classes ficou restrita aos atos relacionados às etapas do processo licitatório de obras públicas, conforme proposto inicialmente pelo trabalho. Nos outros dois cenários buscou-se estender o escopo de forma a aproveitar melhor o conjunto de dados e apresentar resultados mais abrangentes, visto que somente parte da grande quantidade de publicações obtidas se referia a esses tipos de atos.

Relacionando esses cenários com a técnica da validação cruzada e o balanceamento de classes, foram obtidos três conjuntos de dados diferentes para treinamento e teste, onde para cada cenário existem 10 partições diferentes com a divisão de publicações conforme mostrado na Tabela 1.

Tabela 1 – Conjuntos de dados de treinamento e teste resultantes para cada um dos cenários utilizando a técnica de validação cruzada.

Número de Classes	Treinamento	Teste	Total
5 classes	357.300	39.700	397.000
12 classes	86.292	9.588	95.880
20 classes	14.940	1.660	16.600

Tabela produzida pelo autor.

5.2 Resultados Relacionados à Classificação

Todos os modelos utilizados foram testados para diferentes combinações de parâmetros e configurações, fazendo com que sejam adaptados da melhor forma ao problema de classificação sendo resolvido por este trabalho. Na Tabela 2 são mostrados os parâmetros e as configurações finais definidos para cada um dos modelos utilizados, onde na maior parte são utilizadas configurações padrão e definidas as penalidades.

Tabela 2 – Parâmetros e configurações finais para cada modelo.

Modelo	**Configurações e Parâmetros**
kNN	n_neighbors=10
LinearSVC	dual=False, tol=1e-3 penalty='l2'
LinearSVC-L1	dual=False, tol=1e-3, penalty='l1'
LinearSVC-L2	dual=False, tol=1e-3, penalty='l2'
NearestCentroid	-
MultinomialNB	alpha=.01
BernoulliNB	alpha=.01
ComplementNB	alpha=.1
PassiveAggressive	-
RandomForest	max_iter=50
Ridge	tol=1e-2, solver='auto'
SGD-EN	alpha=.0001, max_iter=50, penalty='elasticnet'
SGD-L1	alpha=.0001, max_iter=50, penalty='l1'
SGD-L2	alpha=.0001, max_iter=50, penalty='l2'

Tabela produzida pelo autor.

Assim sendo, foi simulado um processo de classificação real, no qual cada modelo foi testado para cada uma das partições criadas com a técnica de validação cruzada, e treinado como todas as outras, resultando em dez processos de classificação diferentes. Por ser multiclasse, as métricas foram resultantes do cálculo de médias *macro* entre as classes. Já entre as partições, foram consideradas as métricas encontradas no melhor caso.

A abordagem utilizada na avaliação dos resultados foi o de melhor caso, pois foram encontradas dificuldades no cálculo de matrizes de confusão e curvas ROC a partir da média dos valores de verdadeiros/falsos positivos/negativos de todos os procedimentos como um todo, tendo então sido considerado mais vantajoso a escolha dessa abordagem. Isso não foi considerado prejudicial aos resultados deste trabalho, visto que os valores de métricas calculados a partir de cada partição da validação cruzada foram bem semelhantes, não tendo uma queda muito abrupta do melhor para o pior caso.

Os resultados para cada uma dessas partições foram comparados de forma a encontrar o melhor caso geral para todos os modelos, sendo isso feito para cada um dos três cenários. As métricas finais encontradas para o melhor caso são mostradas a seguir, começando pela classificação com cinco classes e terminando com a de vinte classes.

As métricas resultantes (*macro averaging*) do processo de classificação com cinco classes para cada um dos modelos são mostradas na Tabela 3.

Tabela 3 – Métricas resultantes do processo de classificação com cinco classes para cada modelo.

Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-*Score***
kNN	0.9326	0.9327	0.9326	0.9325
LinearSVC	0.9769	0.9769	0.9769	0.9768
LinearSVC-L1	0.9772	0.9773	0.9772	0.9772
LinearSVC-L2	**0.9788**	**0.9789**	**0.9788**	**0.9788**
NearestCentroid	0.8265	0.8427	0.8265	0.8283
MultinomialNB	0.9134	0.9151	0.9134	0.9129
BernoulliNB	0.8885	0.8936	0.8885	0.8854
ComplementNB	0.9099	0.9116	0.9099	0.9090
PassiveAggressive	0.9785	0.9785	0.9785	0.9785
RandomForest	0.9642	0.9649	0.9642	0.9640
Ridge	0.9752	0.9754	0.9752	0.9752
SGD-EN	0.9407	0.9428	0.9407	0.9401
SGD-L1	0.9361	0.9377	0.9361	0.9355
SGD-L2	0.9450	0.9468	0.9450	0.9444

Tabela produzida pelo autor.

A Figura 17 mostra a matriz de confusão resultante da classificação com cinco classes para o modelo *LinearSVC-L2*, que foi destacado na Tabela 3 como o de melhor desempenho.

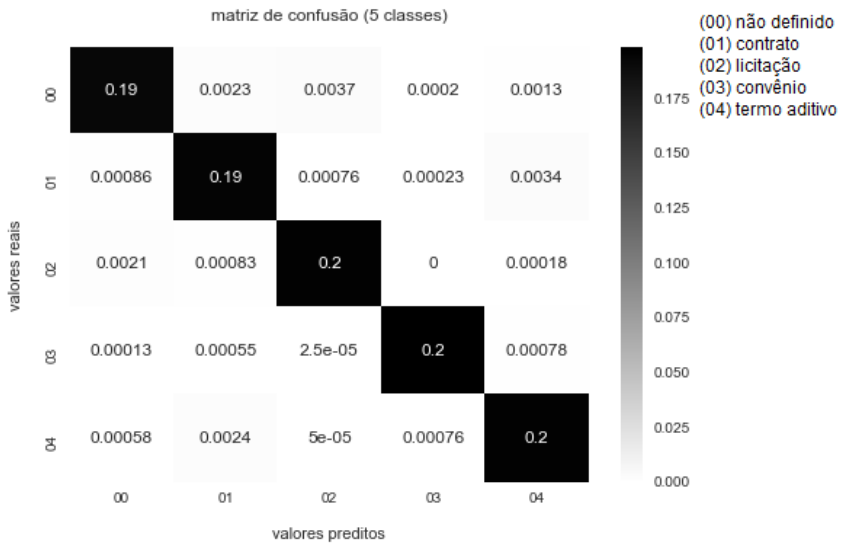

Figura 17 – Matriz de confusão para o modelo *LinearSVC-L2* (classificação com cinco classes).

Figura produzida pelo autor.

As métricas resultantes (*macro averaging*) do processo de classificação com doze classes para cada um dos modelos são mostradas na Tabela 4.

Tabela 4 – Métricas resultantes do processo de classificação com doze classes para cada modelo.

Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-*Score***
kNN	0.4205	0.9240	0.4205	0.4555
LinearSVC	0.9754	0.9754	0.9754	0.9754
LinearSVC-L1	0.9752	0.9753	0.9752	0.9752
LinearSVC-L2	0.9759	0.9759	0.9759	0.9759
NearestCentroid	0.8603	0.8796	0.8603	0.8655
MultinomialNB	0.9254	0.9262	0.9254	0.9255
BernoulliNB	0.8464	0.8603	0.8464	0.8424
ComplementNB	0.9231	0.9244	0.9231	0.9216
PassiveAggressive	0.9749	0.9749	0.9749	0.9749
RandomForest	0.9588	0.9588	0.9588	0.9587
Ridge	**0.9760**	**0.9761**	**0.9760**	**0.9760**
SGD-EN	0.9549	0.9550	0.9549	0.9548
SGD-L1	0.9473	0.9472	0.9473	0.9471
SGD-L2	0.9604	0.9604	0.9604	0.9602

Tabela produzida pelo autor.

A Figura 18 mostra a matriz de confusão resultante da classificação com doze classes para o modelo *Ridge*, que foi destacado na Tabela 4 como o de melhor desempenho.

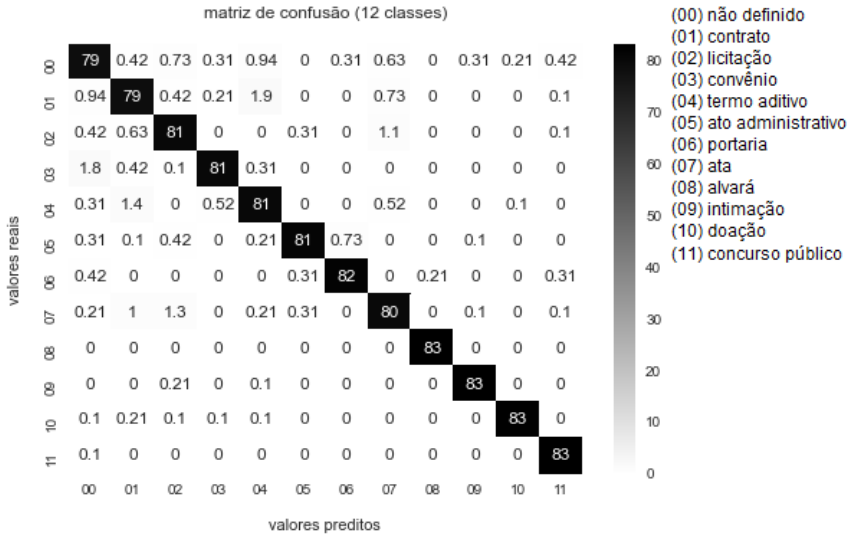

Figura 18 – Matriz de confusão para o modelo *Ridge* (classificação com doze classes).

Figura produzida pelo autor.

As métricas resultantes (*macro averaging*) do processo de classificação com 20 classes para cada um dos modelos são mostradas na Tabela 5.

Tabela 5 – Métricas resultantes do processo de classificação com vinte classes para cada modelo.

Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-*Score***
kNN	0.3259	0.8403	0.3259	0.3731
LinearSVC	0.9578	0.9586	0.9578	0.9575
LinearSVC-L1	0.9506	0.9510	0.9506	0.9501
LinearSVC-L2	0.9590	0.9595	0.9590	0.9587
NearestCentroid	0.8512	0.8809	0.8512	0.8604
MultinomialNB	0.8898	0.8959	0.8898	0.8898
BernoulliNB	0.8127	0.8388	0.8127	0.8083
ComplementNB	0.8861	0.8877	0.8861	0.8808
PassiveAggressive	**0.9627**	**0.9632**	**0.9627**	**0.9625**
RandomForest	0.9301	0.9318	0.9301	0.9286
Ridge	0.9554	0.9562	0.9554	0.9548
SGD-EN	0.9404	0.9407	0.9404	0.9383
SGD-L1	0.9295	0.9296	0.9295	0.9271
SGD-L2	0.9512	0.9516	0.9512	0.9499

Tabela produzida pelo autor.

A Figura 19 mostra a matriz de confusão resultante da classificação com 20 classes para o modelo *PassiveAgressive*, que foi destacado na Tabela 5 como o de melhor desempenho.

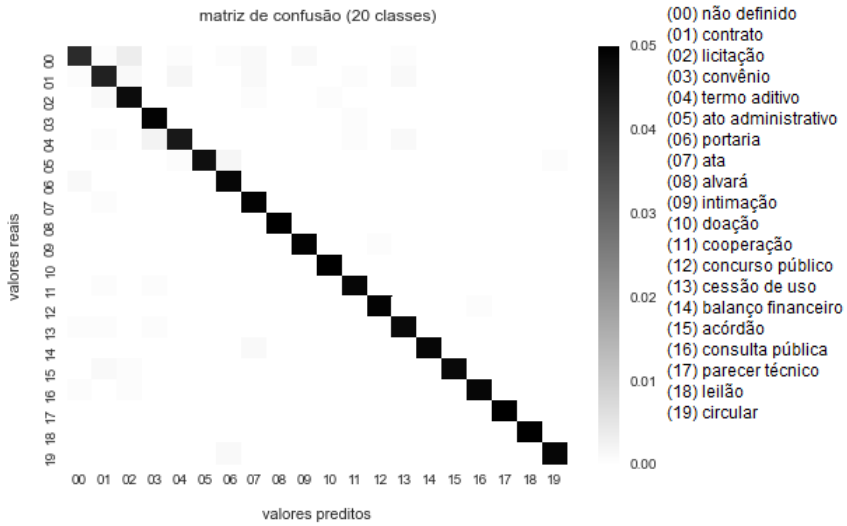

Figura 19 – Matriz de confusão para o modelo *PassiveAgressive* (classificação com vinte classes).

Figura produzida pelo autor. Sem os valores numéricos por falta de espaço.

5.3 Análise dos Resultados

Os resultados mostram que quase todos os modelos utilizados obtiveram um ótimo desempenho na resolução da tarefa, tendo somente o *kNN* decaído muito com o aumento do número de classes. No geral, o desempenho de todos os modelos decaiu um pouco com a adição de novas classes e, consequente, redução do número de publicações, sendo isso por um motivo ou pelo outro.

Além do desempenho extremamente baixo do *kNN* para os cenários com doze e vinte classes, chegando a valores próximos da faixa de 0.3, os modelos baseados no Teorema de Bayes tiveram uma performance um pouco abaixo das demais, ficando próximos da faixa de 0.9 para todos os cenários.

Assim sendo obteve-se como destaque os modelos *SVCLinear-L2*, *Ridge* e *PassiveAgressive*, que atingiram bons resultados em todos os cenários, chegando até a valores bem próximos da faixa de 0.98, que estão acima do que era esperado inicialmente. O melhor modelo no cenário com cinco classes foi o *SVCLinear-L2*, que obteve um F1-*Score* de 0.9788, já no de doze classes foi o *Ridge*, com um F1-*Score* de 0.9760, e no caso com vinte classes foi o *PassiveAgressive*, com um F1-*Score* de 0.962.

A partir das taxas de verdadeiros positivos e falsos positivos resultantes do melhor caso de classificação, foram calculadas curvas ROC para o melhor modelo em cada cenário, sendo essas curvas mostradas a seguir.

A Figura 20 mostra a curva ROC resultante do processo de classificação com cinco classes utilizando o modelo *SVCLinear-L2*.

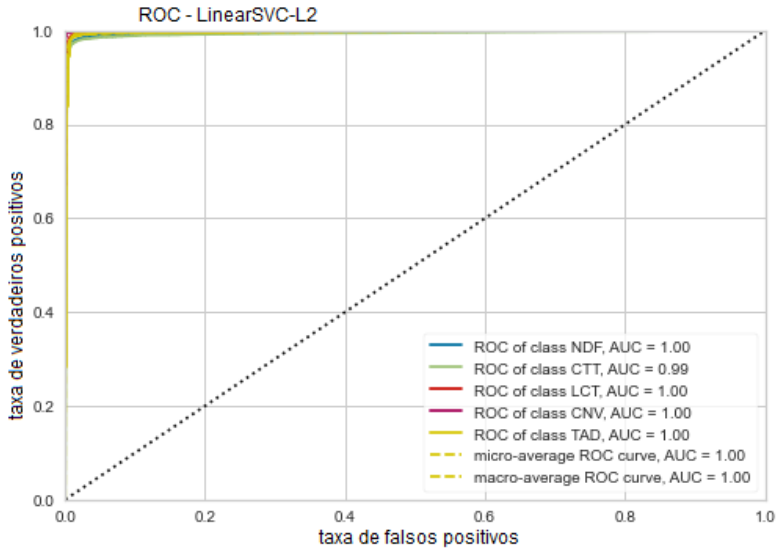

Figura 20 – Curva ROC resultante da classificação para o modelo *SVCLinear-L2* (5 classes).

A Figura 21 mostra a curva ROC resultante do processo de classificação com doze classes utilizando o modelo *Ridge*.

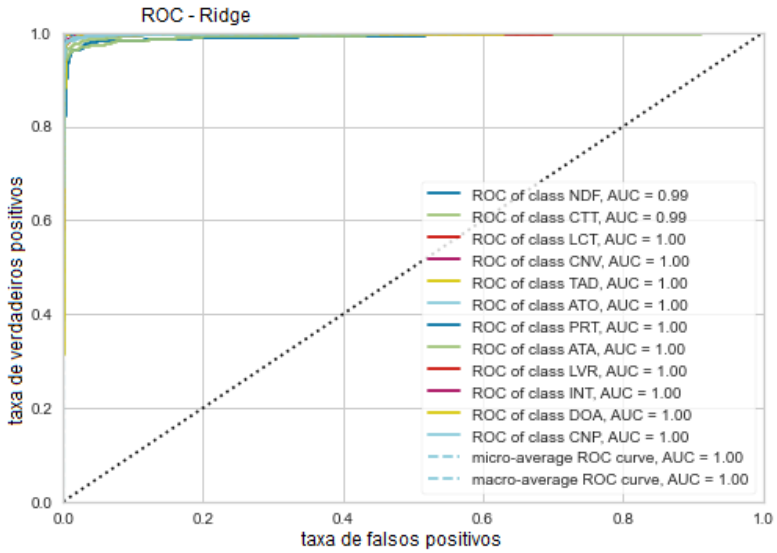

Figura 21 – Curva ROC resultante da classificação para o modelo *Ridge* (12 classes).

A Figura 22 mostra a curva ROC resultante do processo de classificação com vinte classes utilizando o modelo *Passive Agressive*.

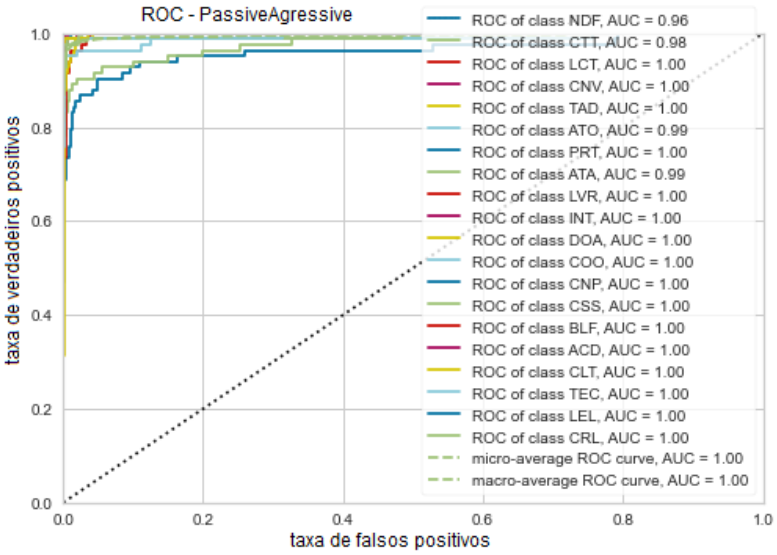

Figura 22 – Curva ROC resultante da classificação para o modelo *Passive Agressive* (20 classes).

Por meio de uma análise qualitativa as curvas ROC deixam claro que os resultados foram muito bons, visto que para todos os três cenários houve maximização da taxa de verdadeiros positivos e minimização da taxa de falsos positivos, mostrando que os modelos são bons em acertar quando uma publicação faz parte de uma categoria. Essas curvas mostram também que o que distanciou os resultados da perfeição foram os verdadeiros negativos e falsos negativos, que consistem no erro dos modelos ao dizerem que uma publicação não faz parte daquela categoria. Isso pode ser relacionado à atribuições de publicações à classe 00 (não definido), visto que essa classe acabou ficando bem abrangente, além de poder possuir algum ruído, que é referente a um pequeno número de publicações de outras classes.

6 Conclusão e Trabalhos Futuros

Neste capítulo é feito um apanhado geral sobre tudo que foi apresentado ao longo do desenvolvimento deste trabalho, relacionando o que foi proposto inicialmente com os resultados finais obtidos. São também apontadas direções para a realização de possíveis trabalhos no futuro, sejam eles melhorias ou extensões do que foi feito aqui, ou que tenham pelo menos alguma conexão com os temas aqui abordados.

6.1 Conclusão

O proposto inicialmente como objetivo principal deste trabalho foi a criação de uma metodologia de classificação de publicações informativas relacionadas a processos licitatórios de obras públicas disponíveis na Seção 3 do Diário Oficial da União. Essa metodologia teria que produzir um classificador que não só funcionasse de forma automática, mas que também mantivesse o nível de qualidade do que seria o trabalho de um especialista. Para que isso fosse possível, foram selecionados 14 modelos de classificação diferentes, onde todos tiveram seu desempenho avaliado estatisticamente por meio do cálculo da acurácia, precisão, sensibilidade e F1-*score*, obtidas durante testes que simulavam um processo de classificação real. Todos esses modelos tinham seu funcionamento baseado no uso de aprendizado de máquina supervisionada, e os dados utilizados como entrada foram tratados com técnicas de processamento de linguagem natural. Isso foi feito para três cenários diferentes: 5, 12 e 20 classes, e por mais que o desempenho dos modelos em geral tenha decaído um pouco à medida que o número de classes aumentasse e o de publicações por classe diminuísse, ainda foram obtidos resultados satisfatórios para todos os cenários. Os modelos *SVCLinear-L2*, *Ridge* e *PassiveAgressive* se sobressairam na tarefa, apresentando respectivamente os valores de **0.9788**, 0.9752 e 0.9785 para o F1-*score* no melhor caso da validação cruzada para a classificação com cinco classes, 0.9759, **0.9760** e 0.9749 com doze classes, e 0.9587, 0.9548 e **0.9625** com vinte classes.

Também foram propostos alguns objetivos secundários: a criação de um novo conjunto de dados, a avaliação de diferentes métodos de classificação na resolução do problema variando configurações e parâmetros, a comparação entre os resultados utilizando diferentes técnicas de pré-processamento e a contribuição ao avanço do projeto DV. Os três primeiros foram também alcançados e fizeram parte do processo de criação da metodologia de classificação, enquanto o último foi atingido como consequência dos demais.

Espera-se que a metodologia de classificação desenvolvida sirva como uma forma de organizar e controlar o grande fluxo de dados gerados pelos processos licitatórios de obras públicas e ajude no trabalho da PF na busca por irregularidades, produzindo indiretamente um impacto positivo na vida das pessoas a longo prazo.

Por fim, pode-se confirmar a partir dos bons resultados apresentados que o uso de aprendizado de máquina em conjunto com técnicas de processamento de linguagem natural são suficientemente capazes de resolver o problema apontado pela hipótese de pesquisa, obtendo resultados equiparáveis aos de um ser humano capacitado. Isso abre caminho para novos estudos visando avaliar a possibilidade de resolução de outros problemas semelhantes e, possivelmente, mais complexos.

6.2 Trabalhos Futuros

Os resultados apresentados, apesar de terem sido bastante satisfatórios dentro do que foi inicialmente proposto, foram bem limitados em alguns sentidos, abrindo muita margem para melhora e extensão. Isso é evidente quando percebe-se que foram utilizadas apenas publicações feitas no DOU e que existem outros jornais, tanto estaduais quanto municipais que exercem papéis semelhantes. Também é perceptível que vários tipos de publicações não foram utilizadas como classes. Abaixo são listados possíveis caminhos para melhorias:

- Aumentar a quantidade de dados de tipos de publicações com valores baixos e limitantes, possibilitando a criação de novas classes e causando melhora nos resultados;
- Fazer uma adaptação para que publicações feitas em outros diários sejam incluídas, possibilitando o aumento do alcance e da utilidade do trabalho;
- Usar modelos de classificação de maior complexidade, como as redes neurais profundas, possibilitando a melhora dos resultados e a resolução de problemas mais difíceis;
- Fazer testes com novos métodos de pré-processamento e com novas formas de extração de características, e também com combinações diferentes de ambos.

Cada etapa do processo licitatório contém informações específicas que seriam relevantes em uma possível investigação policial, informações essas diretamente ligadas ao tipo de publicação. Uma possibilidade de extensão para o trabalho seria a busca e extração de entidades que constituam essas informações. Se a partir dessa extração fosse possível criar uma forma de relacionar tais entidades entre si, daria para traçar uma espécie de grafo conectando todas as informações relevantes referentes a um dado processo licitatório, o que facilitaria muito o trabalho dos investigadores.

Referências

AFENDRAS, G.; MARKATOU, M. Optimality of training/test size and resampling effectiveness in cross-validation. **Journal of Statistical Planning and Inference**, Elsevier, v. 199, p. 286–301, 2019. Citado nas pp. 31, 32.

AGUIAR, A.; SILVEIRA, R.; PINHEIRO, V.; FURTADO, V.; NETO, J. A. Text Classification in Legal Documents Extracted from Lawsuits in Brazilian Courts. In: SPRINGER. BRAZILIAN Conference on Intelligent Systems. 2021. P. 586–600. Citado na p. 41.

AMPADU, H. **Random Forests Understanding**. 2021. Disponível em: <https://ai-pool.com/a/s/random-forests-understanding>. Acesso em: 21 abr. 2021. Citado na p. 30.

AZEVEDO, F. Corrupção, mıdia e escândalos midiáticos no Brasil. **Debate, Belo Horizonte**, v. 2, n. 3, p. 14–19, 2010. Citado na p. 12.

BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS journal of photogrammetry and remote sensing**, Elsevier, v. 114, p. 24–31, 2016. Citado na p. 29.

BOCHENEK, A. C.; PEREIRA, J. L. Corrupção sistêmica no Brasil: Enfrentamento e dificuldades. **Revista Jurıdica da FANAP**, v. 5, n. 1, 2018. Citado na p. 12.

BOTTOU, L. Stochastic gradient descent tricks. In: NEURAL networks: Tricks of the trade. Springer, 2012. P. 421–436. Citado na p. 31.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern recognition**, Elsevier, v. 30, n. 7, p. 1145–1159, 1997. Citado na p. 37.

BRASIL. **Decreto nº 9.215, de 29 de novembro de 2017**. 2017. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/D9215.htm. Acesso em: 9 abr. 2021. Citado na p. 20.

BRASIL. **Imprensa nacional**. 2022. http://www.in.gov.br. Acesso em: 9 abr. 2022. Citado nas pp. 14, 20.

BRASIL. **Lei nº 10.520, de 17 de julho de 2002**. 2002. http://www.planalto.gov.br/ccivil_03/LEIS/2002/L10520.htm. Acesso em: 9 abr. 2021. Citado na p. 17.

BRASIL. **Lei nº 14.133, de 1 de abril de 2021**. 2021. http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm. Acesso em: 9 abr. 2021. Citado na p. 18.

BRASIL. **Lei nº 8.666, de 21 de junho de 1993**. 1993. http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm. Acesso em: 9 abr. 2021. Citado nas pp. 14, 17.

BROWNE, M. W. Cross-Validation Methods. **Journal of Mathematical Psychology**, v. 44, n. 1, p. 108–132, 2000. ISSN 0022-2496. DOI: https://doi.org/10.1006/jmps.1999.1279. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022249699912798>. Citado na p. 32.

CADE. **Combate a cartéis em licitação**. Conselho Administrativo de Defesa Econômica, 2019. Citado na p. 13.

CASTRO, F. d. A corrupção no orçamento: fraudes em licitações e contratos com o emprego de empresas inidôneas. **Artigo apresentado ao Instituto Serzedello Corrêa–ISC/TCU para a obtenção do tıtulo de Especialista em Orçamento Público. Brasılia: TCU**, 2010. Citado na p. 13.

CHALKIDIS, I.; FERGADIOTIS, M.; MALAKASIOTIS, P.; ANDROUTSOPOULOS, I. Large-scale multi-label text classification on EU legislation. **arXiv preprint arXiv:1906.02192**, 2019. Citado na p. 40.

CHOWDHARY, K. Natural language processing. **Fundamentals of artificial intelligence**, Springer, p. 603–649, 2020. Citado na p. 13.

CIOCCARI, D. Operação Lava Jato: escândalo, agendamento e enquadramento. **Revista Alterjor**, v. 12, n. 2, p. 58–78, 2015. Citado na p. 12.

CRAMMER, K.; DEKEL, O.; KESHET, J.; SHALEV-SHWARTZ, S.; SINGER, Y. Online passive aggressive algorithms, 2006. Citado na p. 29.

DE ARAUJO, P. H. L.; CAMPOS, T. E. de; BRAZ, F. A.; SILVA, N. C. da. VICTOR: a dataset for Brazilian legal documents classification. In: PROCEEDINGS of the 12th Language Resources and Evaluation Conference. 2020. P. 1449–1458. Citado na p. 40.

E-DOU. **O Que é o Diário Oficial da União**. Disponível em: <https://e-dou.com.br/o-que-e-o-diario-oficial-da-uniao/>. Acesso em: 20 abr. 2021. Citado na p. 19.

GANTZ, J.; REINSEL, D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. **IDC iView: IDC Analyze the future**, v. 2007, n. 2012, p. 1–16, 2012. Citado na p. 13.

GARDNER, W. A. Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. **Signal processing**, Elsevier, v. 6, n. 2, p. 113–133, 1984. Citado na p. 31.

GOLDBERG, Y. Neural network methods for natural language processing. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 10, n. 1, p. 1–309, 2017. Citado na p. 21.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. **arXiv preprint arXiv:2008.05756**, 2020. Citado nas pp. 33, 35, 36.

JIANG, S.; PANG, G.; WU, M.; KUANG, L. An improved K-nearest-neighbor algorithm for text categorization. **Expert Systems with Applications**, v. 39, n. 1, p. 1503–1509, 2012. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2011.08.040. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417411011511>. Citado na p. 25.

KOWSARI, K.; JAFARI MEIMANDI, K.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text Classification Algorithms: A Survey. **Information**, v. 10, n. 4, 2019. ISSN 2078-2489. DOI: 10.3390/info10040150. Disponível em: <https://www.mdpi.com/2078-2489/10/4/150>. Citado nas pp. 22–24, 26–28, 40.

LIMA, M. C. Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras. pt, p. 126, 2021. Citado nas pp. 38, 39, 62.

MCINTYRE, R. M.; BLASHFIELD, R. K. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. **Multivariate Behavioral Research**, Taylor & Francis, v. 15, n. 2, p. 225–238, 1980. Citado na p. 27.

MOREIRA, E. B.; GUIMARÃES, F. V. Licitação pública. **São Paulo: Malheiros**, 2012. Citado na p. 17.

NOBLE, W. S. What is a support vector machine? **Nature biotechnology**, Nature Publishing Group, v. 24, n. 12, p. 1565–1567, 2006. Citado na p. 26.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na p. 45.

POOLSAWAD, N.; KAMBHAMPATI, C.; CLELAND, J. Balancing class for performance of classification with a clinical dataset. In: PROCEEDINGS of the World Congress on Engineering. 2014. v. 1, p. 1–6. Citado na p. 33.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: CITESEER, 1. PROCEEDINGS of the first instructional conference on machine learning. 2003. v. 242, p. 29–48. Citado na p. 24.

RIFKIN, R. M.; LIPPERT, R. A. Notes on regularized least squares, 2007. Citado na p. 30.

SARITAS, M. M.; YASAR, A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. **International journal of intelligent systems and applications in engineering**, v. 7, n. 2, p. 88–91, 2019. Citado na p. 28.

SCIKIT-LEARN, M. L. i. P. **Receiver operating characteristic (ROC)**. Disponível em: <https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html>. Citado na p. 37.

SINGH, G.; KUMAR, B.; GAUR, L.; TYAGI, A. Comparison between multinomial and Bernoulli naıve Bayes for text classification. In: IEEE. 2019 International Conference on Automation, Computational and Technology Management (ICACTM). 2019. P. 593–596. Citado na p. 40.

VAJJALA, S.; MAJUMDER, B.; GUPTA, A.; SURANA, H. **Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems**. O'Reilly Media, 2020. ISBN 9781492054054. Disponível em: <https://books.google.com.br/books?id=G40jywEACAAJ>. Citado na p. 21.

VALLIM, J. J. d. C. B. Uso do modelo de raciocınio baseado em casos para monitoramento de conluio em licitações de obras de pavimentação urbana, 2020. Citado na p. 40.

VAN DER MAATEN, L.; POSTMA, E.; VAN DEN HERIK, J. et al. Dimensionality reduction: a comparative. **J Mach Learn Res**, v. 10, n. 66-71, p. 13, 2009. Citado na p. 25.

VIJAYARANI, S.; ILAMATHI, M. J.; NITHYA, M. et al. Preprocessing techniques for text mining-an overview. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2015. Citado na p. 23.

Anexos

ANEXO A – Exemplo de Formatação dos Arquivos de Publicações Obtidos

Campo	Valor
id	13912578
name	EXTRATO CONTRATO - INFRA - EM 30
idOficio	5793652
pubName	DO3
artType	Extrato de Contrato
pubDate	31/03/2020
artClass	00016:00131:00053:00000:00000:00000:00000:00000: 00000:00000:00032:00000
artCategory	Ministério da Educação/Fundação Universidade de Brasília/- Secretaria de Infraestrutura
artSize	12
artNotes	
numberPage	63
pdfPage	http://pesquisa.in.gov.br/imprensa/jsp/visualiza/ index.jsp?data=31/03/2020&jornal=530&pagina=63
editionNumber	62
highlightType	
highlightPriority	
highlight	
highlightimage	
highlightimagename	
idMateria	12542436
Identifica	EXTRATO DE INSTRUMENTO CONTRATUAl
Campo	**Valor**
text	EXTRATO DE INSTRUMENTO CONTRATUAl <p class=identifica > EXTRATO DE INSTRUMENTO CONTRATUAl Espécie: Contrato nº 1106/2020 Nº Processo: 23106. 056915/ 2016- 62. Regime Diferenciado de Contratação nº 17/2019 - INFRA/UnB Contratante: UNIVERSIDADE DE BRASÍLIA, CNPJ 00.038.174/0001-43 Contratado: CMP CONSTRUTORA MARCELINO PORTO EIRELI, CNPJ 38. 027. 876/ 0001- 02&8203; Objeto: Obra de reforma para o Laboratório e Acervo de Fósseis, Minerais e Rochas, localizado no Campus UnB Planaltina (FUP) da Universidade de Brasília, Planaltina/DF Fundamento legal: Lei nº 8.666/93 e suas alterações, Lei nº 10.406/02, Lei nº 12.462/2011, e Decreto nº 7.581/2011. Vigência: 11/03/2020 a 11/08/2020Valor Global: R$ 44.500,00 Fonte: 2020NE800260.Data de Assinatura: 11/03/2020 Espécie: Primeiro termo aditivo ao contrato nº 1112/2019 Nº Processo: 23106. 018 699 / 2017-38 Contratante: UNIVERSIDADE DE BRASÍLIA, CNPJ 00.038.174/0001-43 Contratado: 3R ENGENHARIA EIRELI, CNPJ 07.371.427/0001-45; Objeto: Prorrogação do prazo de vigência por mais 60 (sessenta) dias e do prazo de execução por mais 60 (sessenta) dias do Contrato nº 1112/2019 INFRA/UNB.Fundamento legal: Art. 57, § 1º , incisos II da Lei 8666/93.Vigência: 13/06/2020 a 11/08/2020Data de Assinatura: 1903/2020
date	2020-03-31
_id	2020030050929
indice	15183897

Fonte: Lima (2021, p. 58, 59).