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

Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras

Marcos Cavalcanti Lima

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Orientador
Prof. Dr. Flávio de Barros Vidal

Brasília
2021

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Mestrado em Informática

Coordenador: Prof. Dr. Genaína Nunes Rodrigues

Banca examinadora composta por:

Prof. Dr. Flávio de Barros Vidal (Orientador) — CIC/UnB
Prof. Dr. Ana Carolina Lorena — ITA
Prof. Dr. Luís Paulo Faina Garcia — PPGI/CIC/UnB
Prof. Dr. Camilo Chang Dórea (Suplente) — PPGI/CIC/UnB

CIP — Catalogação Internacional na Publicação

Lima, Marcos Cavalcanti.

Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras / Marcos Cavalcanti Lima. Brasília : UnB, 2021.
126 p. : il. ; 29,5 cm.

Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2021.

1. Reconhecimento de Padrões, 2. Processamento Linguagem Natural, 3. Licitações de Obras Públicas, 4. Conluio

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

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

Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras

Marcos Cavalcanti Lima

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

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

Prof. Dr. Ana Carolina Lorena
ITA

Prof. Dr. Luís Paulo Faina Garcia
PPGI/CIC/UnB

Prof. Dr. Camilo Chang Dórea (Suplente)
PPGI/CIC/UnB

Prof. Dr. Genaína Nunes Rodrigues
Coordenador do Mestrado em Informática

Brasília, 31 de Maio de 2021

Dedicatória

À minha amada esposa e nossos amados filhos.

Agradecimentos

Ao Diretor do Instituto Nacional de Criminalística Perito Criminal Federal Perito Criminal Federal Raimundo Nonato de Azevedo Filho e Diretor Técnico-Científico Perito Criminal Federal Alan de Oliveira Lopes por incentivar e apoiar a realização desta pesquisa.

Ao ex-Diretor do Instituto Nacional de Criminalística Perito Criminal Federal Luiz Spricigo Junior e ao ex-Diretor Técnico-Científico Perito Criminal Federal Fábio Augusto da Silva Salvador que se envolveram e se empenharam na realização deste projeto.

Aos colegas do Serviço de Perícias em Engenharia que me apoiaram direta e indiretamente.

Ao Professor Dr. Flávio de Barros Vidal que desde os primeiros momentos entendeu as demandas e contribuiu para superação dos desafios das pesquisas da criminalística da Polícia Federal.

Este trabalho foi desenvolvido no âmbito do convênio celebrado entre a Polícia Federal e a Universidade de Brasília no projeto "Pesquisa Aplicada de Inovações Tecnológicas no domínio da Perícia Criminal Federal".

Resumo

A identificação de fraudes e conluios em licitações de obras públicas é uma tarefa manual dispendiosa dependente tanto de experiência profissional quanto de profundo conhecimento técnico e legal. As bases de dados públicas, aliadas a dados de licitações e contratos previamente analisados por peritos criminais altamente capacitados, formaram a base de dados passível de ser analisada para a identificação de atos ilícitos. Neste trabalho é proposta uma metodologia para realizar a detecção e classificação automática de padrões de conluio em licitações públicas, utilizando como fontes os dados disponíveis nos principais repositórios oficiais públicos, agregando a utilização de técnicas de reconhecimento de padrões para a realização deste objetivo proposto. Em uma abordagem inicial, obteve-se com sucesso para a formação da base de dados do trabalho um total de 15.132.968 publicações da Seção 3 do Diário Oficial da União em formato de texto e 1.907 documentos como referência de indicativo de atividades de conluio (estes disponibilizados por instituição parceira) que indicavam risco no processo licitatório. Foram testados modelos lineares clássicos, redes neurais profundas, *bottleneck*, Bi-LSTM e multicanal com vetorização do texto com TF-IDF e DOC2VEC, e dados estruturados extraídos do texto. O melhor F1-*score* foi obtido com o modelo *passive-aggressive* com 93,4% e o modelo *bottleneck* obteve 93,0% com melhor precisão.

Palavras-chave: Reconhecimento de Padrões, Processamento Linguagem Natural, Licitações de Obras Públicas, Conluio

Abstract

Identifying fraud and collusion in public bids is an expensive manual task and dependent on professional experience using in-depth technical and legal knowledge. Public databases, allied to bidding and contract data previously analyzed by highly trained criminal experts, form the database that can be analyzed for irregularities identification. This work proposes a methodology for automatic detection and classification of collusion patterns in public bids text, using data sources available on main public official repositories and adding pattern recognition techniques to achieve a model that detects and classifies this pattern. In an initial approach, a total of $15,132,968$ publications of the Diario Oficial da União news, Section 3, in text format and $1,907$ documents as a reference for collusion activities were successfully obtained for the formation of the central work database (provided by a partner institution) that indicated risk in the bidding process. Classic linear models, deep neural networks, bottleneck, Bi-LSTM, and multichannel were tested with text vectorization with TF-IDF and DOC2VEC, and structured data extracted from the text. The best F1-score was obtained with a passive-aggressive model with 93.4%, but the bottleneck model obtained 93.0% with better precision.

Keywords: Pattern Recognition, Natural Language Processing, Public bidding process, Collusion

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

Redes Neurais

a Função de ativação aplicada à saída de uma camada

BL BiLSTM

D Densa

dr Taxa de *dropout*

F Fusão de redes neurais em redes multicanais

FL Achatamento ou *flattener*

op Operação de fusão (F) de redes neurais multicanais

TD Densa distribuída no tempo

u Quantidade de nós de uma camada

Siglas

API *Application Programming Interface* ou Interface de Programação de Aplicação

AUC *Area Under Curve*

BERT *Bidirectional Encoder Representations from Transformers*

BiLSTM *Bidirectional Long-Short-Term Memory*

BOW *Bag of Words*

CADE Conselho Administrativo de Defesa Econômica

CBR *Case-based reasoning*

CGU Controladoria Geral da União

CNPJ Cadastro Nacional da Pessoa Jurídica

CPF Cadastro de Pessoa Física

DNN *Deep Neural Networks*

EER *Equal Error Rate*

HTML Hyper Text Markup Language

IP *Internet Protocol*

JSON *JavaScript Object Notation*

LOD Linking Open Data

LSTM *Long-Short-Term Memory*

NLP Natural Language Processing

OCDE Organização para a Cooperação e Desenvolvimento Econômico

PDF *Portable Document File*

RegEx *Regular Expression* - Expressões regulares

RNN *Recurrent Neural Network*

ROC *Receiver Operating Characteristic*

SAG *Stochastic Average Gradient*

SGD *Stochastic Gradient Descent*

SINAPI Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil

SMOTE *Synthetic Minority Oversampling Technique*

SVM *suport vector machines*

TCU Tribunal de Contas da União

TF-IDF Term Frequency - Inverse Document Frequency

ULMFiT *Universal language model fine-tuning*

XML *Extensible Markup Language*

XML *Extensible Markup Language*

Variáveis

α Fator de redução da função *Leaky* ReLU

$\hat{y}$ Valor predito por uma classificador

λ Peso da regularização $R(\Theta)$ da função de perda média $\mathcal{L}$

$\mathcal{L}$ Perda média ou *average loss*

$\sigma(x)$ Função sigmoide

Θ Conjunto de parâmetros do classificador

$\tilde{y}$ Sinal de $\hat{y}$, assume os valores -1 ou $+1$

$\vec{w}$ Vetor de pesos de um modelo

$\vec{x}$ Vetor de medidas (características) de uma palavra w ou *n-gram*

$\vec{z}$ Vetor de entrada da função de ativação

c Classe atribuída ou estimada de um documento ou texto

d Identificador de um documento

d_n Dimensão de um classificador linear

f Portão de esquecimento de uma rede LSTM

g Função de Ativação

$h^{(n)}$ Saída da camada n de um *perceptron*

i Portão de entrada de uma rede LSTM

L Perda ou *loss*

o Portão de saída de uma rede LSTM

$R(\Theta)$ Termo de regularização da função de perda média $\mathcal{L}$

RNN Função recorrente de uma RNN

S Função sigmóide

w	Palavra em um documento d
$W_{jk}^{(i)}$	Peso da rede aplicado ao valor que vai do nó (ou neurônio) j na camada $i-1$ para o nó k na camada i
y	Valor de referência de uma elemento classificado
b	*Tendência ou bias*
t	Número de elementos na base de dados

Aviso Legal

Os dados apresentados, no âmbito deste Mestrado Acadêmico, foram compilados a partir de dados públicos e da base de conhecimento da rede de Peritos Criminais Federais da Polícia Federal.

As licitações, contratações, obras e/ou convênios foram marcados como *risco = 1* neste banco de dados por múltiplos fatores, dentre os quais podem-se incluir indicadores que, isoladamente ou cumulativamente, envolvam data, local, tipo, partes envolvidas e outras informações vinculadas aos processos.

Da inclusão de um processo neste banco, não se pode concluir pela a ocorrência ou ausência de irregularidade administrativa ou penal.

Disclaimer

The data shown in this Academic Master were compiled from public data and processed from the Brazilian's Federal Police Experts network knowledge.

The procurements, contracts, works, and/or agreements were marked as *risk = 1* in this database by multiply reasons that may include singly or cumulatively indicators about date, place, type, parts, and any other information linked to the process.

The process's presence in this database cannot lead to conclusion about legal or criminal status.

1

Introdução

> *"Fazei tudo por Amor. - Assim não há coisas pequenas: tudo é grande. - A perseverança nas pequenas coisas, por Amor, é heroísmo."*
>
> – São Josemaría Escrivá

O combate a fraudes em licitações, principalmente em licitações de obras, se apresenta como um grande desafio para a sociedade ao envolver grandes somas em dinheiro e alta complexidade técnica e burocrática. Além disso, o desvio de recursos, objetivo destas fraudes, é precedido de ações que visam direcionar ou manipular o resultado das licitações [35]. Por essa razão, inicialmente, é preciso conceituar outros temas não afetos à pesquisa computacional. O CADE (Conselho Administrativo de Defesa Econômica) define Cartel em sua cartilha Combate a Cartéis em Licitações [35] da seguinte forma:

> (...) Cartel é um acordo explícito ou implícito entre concorrentes para, principalmente, fixação de preços ou quotas de produção, divisão de clientes e de mercados de atuação. Cartéis são considerados a mais grave lesão à concorrência porque prejudicam seriamente os consumidores ao aumentar preços e restringir a oferta, tornando os bens e serviços mais caros ou indisponíveis (...)

De fato, os processos licitatórios de obras públicas geram uma enorme gama de documentos, planilhas e projetos, tornando crítica a determinação e identificação de casos de cartel em obras públicas diante da quantidade de informação gerada e não processada. Apenas uma obra pode apresentar dezenas de publicações (objeto de análise desta pesquisa) de procedimentos (convênios, licitação, contratos e aditivos) em diário oficial.

Vários órgãos públicos possuem em seu escopo a atividade de fiscalização e repressão destes desvios, tais como Ministério Público, Justiça Federal, órgãos de controle (Tribunal de Contas da União - TCU, atribuição dada pelo artigo 71 da Constituição Federal [24], Controladoria Geral da União - CGU, atribuição dada pela Lei Federal 10.683 [19] e equivalentes estaduais e municipais), Polícias Civis e a Polícia Federal (atribuição dada pelo Art. 144 da Constituição Federal [24] e Decreto-Lei73.332/73 [16]) .

O desafio de processamento destas informações aumentará, no contexto de uma investigação policial, onde os limites legais restritivos e a lógica do processo criminal tende a levar a uma ação corretiva e não preventiva e, por isso, não contemporânea aos fatos. Assim, a análise da documentação gerada durante o processo é fundamental da determinação da verdade real dos fatos.

1.1 Definição do Problema e Justificativas

O desafio de determinação de todos os tipos de fraudes em processos públicos é mais evidente na detecção de conluio e determinação de cartel. Nesta seção é abordado o problema sob a ótica da investigação policial e como é feita a análise manual e supervisionada, sob a ótica da análise de dados além a utilização dos textos das publicações do Diário Oficial da União como forma de enfrentar este desafio.

1.1.1 Sob a Ótica da Investigação Policial

Nos últimos cinco anos o governo federal do Brasil investiu [30] aproximadamente 283,8 bilhões de reais em 23.352 contratos de obras públicas. Estes contratos abrangiam a mais variada gama de projetos incluindo refinarias, portos, estádios de futebol, usinas de geração de energia, túneis e barragens em um país de dimensões continentais como o Brasil, fazendo das licitações de obras públicas um campo fértil à fraude e ao conluio [97] e, no Brasil, grande oportunidade para a ação fraudulenta e desvio de recursos públicos [58].

A Polícia Federal tem trabalhado em investigações em fraudes em obras públicas nas últimas quatro décadas e desenvolveu a base de sua metodologia investigativa em um grupo especializado de peritos criminais com formação nas áreas de engenharia civil, elétrica e mecânica, além de peritos em informática e contabilidade [8]. Os tipos de fraudes investigadas são principalmente o conluio em licitações e superfaturamento de preço, quantidade e qualidade [89]. Utiliza-se O conhecimento acumulado durante estas décadas para fortalecer o conhecimento sobre os dados analisados nesta pesquisa.

Dentro dos procedimentos de perícias criminais ou auditoria de obras públicas, o processo de determinação de superfaturamento encontra-se bem delimitado e com procedimentos conhecidos e debatidos na comunidade científica da área. No final dos anos 2000, a equipe de peritos da Polícia Federal elaborou um método para o cálculo e classificação do superfaturamento culminado no livro *Superfaturamento de Obras Públicas*, de Lopes [89] e com a inclusão destes conceitos na Lei 14.133/2021[25].

Entretanto, surgem problemas que usualmente limitam as ações dos processos criminais. Primeiro, como comprovado por Silva Filho et al. [117] em 2010, é possível que elementos de organizações criminosas se sagrem vencedoras de licitações públicas por

meios escusos, abrindo inclusive margem para o pagamento de propina, mesmo praticando preços globais abaixo da referência oficial de custos para obras públicas, o SINAPI[61].

Um segundo ponto a considerar é a observação empírica da inefetividade de ações penais que, apesar de demonstrar o superfaturamento de contratos público, falharam em determinar ações de cartelização e/ou conluio no processo licitatório. Ações penais instruídas desta forma tendem a não lograr êxito em alcançar a punição dos agentes responsáveis pelas mais diversas formas de superfaturamento em obras públicas [44].

Um último ponto que pode ser levantado sobre a importância da investigação do conluio é que, como demonstrado por Lima [86] para obras rodoviárias quando são observadas condições de competitividades é possível obter descontos médios da ordem de 37% em relação aos valores de referência, contra 5% em procedimentos sem competitividade e que fora observado [90, 113] que dificilmente é alcançado o superfaturamento dos contratos quando não houve qualquer forma de fraude na licitação.

1.1.2 Análise Manual e Supervisionada

Lopes [90] demonstra o processo manual de identificação e comprovação de um processo de conluio de licitações públicas para a Operação Caixa de Pandora da Polícia Federal seguindo o processo descrito pelo CADE [35] que lista as formas de agir de um cartel:

> **a) Fixação de preços**, na qual há um acordo firmado entre concorrentes para aumentar ou fixar preços e impedir que as propostas fiquem abaixo de um "preço base".
>
> **b) Direcionamento privado da licitação**, em que há a definição de quem irá vencer determinado certame ou uma série de processos licitatórios, bem como as condições nas quais essas licitações serão adjudicadas.
>
> **c) Divisão de mercado**, representada pela divisão de um conjunto de licitações entre membros do cartel, que, assim, deixam de concorrer entre si em cada uma delas. Por exemplo, as empresas A, B e C fazem um acordo pelo qual a empresa A apenas participa de licitações na região Nordeste, a empresa B na região Sul e a empresa C na região Sudeste.
>
> **d) Supressão de propostas**, modalidade na qual concorrentes que eram esperados na licitação não comparecem ou, comparecendo, retiram a proposta formulada, com intuito de favorecer um determinado licitante, previamente escolhido.
>
> **e) Apresentação de propostas "pro forma"**, caracterizada quando alguns concorrentes formulam propostas com preços muito altos para serem aceitos ou entregam propostas com vícios reconhecidamente desclassificatórios. O objetivo dessa conduta é, em regra, direcionar a licitação para um concorrente em especial.
>
> **f) Rodízio**, acordo pelo qual os concorrentes alternam-se entre os vencedores de uma licitação específica. Por exemplo, as empresas A, B e C combinam que a primeira licitação será vencida pela empresa A, a segunda pela empresa B, a terceira pela empresa C e assim sucessivamente.
>
> **g) Sub-contratação**, pela qual concorrentes não participam das licitações ou desistem das suas propostas, a fim de serem sub-contratados pelos vencedores. O

> vencedor da licitação a um preço supra-competitivo divide o sobre-preço com o subcontratado.

Um processo burocrático de licitação pública que segue a Lei de Licitações [17] ou a Lei dos Pregões [18] é complexo e envolve a produção de uma quantidade elevada de documentos e informações.

A equipe de perícias deve então analisar toda esta documentação e confrontá-la com informações externas à base de dados e provenientes da investigação para detectar e comprovar que houve conluio ou fraude no processo competitivo da licitação.

1.1.3 Sob a Ótica da Análise de Dados

Como visto anteriormente todos os procedimentos licitatórios geram uma grande quantidade de informação que pode ser interpretada de várias formas diferentes.

Várias iniciativas dos tribunais de contas e de controladorias (estaduais e federais) [10, 105, 116] encararam o desafio de classificação de procedimentos licitatório por levantamento dessas características utilizando-se de dados estruturados extraídos do sistema federal de compras públicas ComprasNet[1]. Os dados estão limitados aos presentes no sistema e baseiam-se na detecção de comportamentos típicos de processos fraudulentos como baixos descontos e baixa competitividades.

Também foi desenvolvida técnica de análise econométrica, como utilizado nas investigações da Lava Jato [113], que, por técnicas estatísticas e econômicas, demonstram a verossimilhança da ação de cartéis.

Outras iniciativas [7, 78, 122, 125] seguiram caminhos semelhantes aos acima descritos, todos eles baseados na utilização de dados estruturados.

É importante salientar que a Lei 8.666 [17] prevê a obrigatoriedade de publicação dos atos oficiais relativos ao processo licitatório em seu artigo 21:

> Art. 21. Os avisos contendo os resumos dos editais das concorrências, das tomadas de preços, dos concursos e dos leilões, embora realizados no local da repartição interessada, deverão ser publicados com antecedência, no mínimo, por uma vez:
>
> I - no **Diário Oficial da União**, quando se tratar de licitação feita por órgão ou entidade da Administração Pública Federal e, ainda, **quando se tratar de obras financiadas parcial ou totalmente com recursos federais ou garantidas por instituições federais**;[grifado]

Assim, as bases de dados dos Diários Oficias tem muitas informações acerca das licitações públicas e possuem dados abertos em quase todo o mundo [67] e para estas bases foram estudados, por exemplo, para problemas de reconhecimento de entidade [6], vinculação de dados abertos em [5] e reconhecimento de entidades [50] mas não para problemas

[1] https://www.comprasgovernamentais.gov.br/

de detecção de fraudes pois, na sua maioria, não trazem dados anotados ou estruturados ou anotados para este fim.

1.1.4 Diário Oficial da União

Em 1º de outubro de 1862, foi criado o jornal oficial do Império do Brasil com o nome de *Diário Official*, Figura 1.1. Já em 2001, quando as edições eletrônicas começaram a ser disponibilizadas no site `www.in.gov.br`, foi adotado o nome atual de Diário Oficial da União com a sigla DOU [29].


DIARIO OFFICIAL.

IMPERIO DO BRASIL.

ANNO DE 1862. QUARTA FEIRA, 1.º DE OUTUBRO. NUMERO 1.

DIARIO OFFICIAL.

PARTE OFFICIAL.

Figura 1.1: Capa da primeira edição do *Diário Official* [29]

A publicação do DOU é regida pelo Decreto nº 9.215/2018 [23] e ele é dividido em três seções:

- **Seção 1**: os atos normativos de interesse geral dos poderes da União;
- **Seção 2**: os atos relativos aos servidores da administração pública federal; e
- **Seção 3**: os atos decorrentes das contratações públicas e outros de particulares determinados pela legislação.

Como o presente trabalho é voltado à análise risco de fraude de licitações e contratos, apenas a Seção 3 será analisada.

1.2 Hipótese de Pesquisa

A hipótese a ser investigada é de que é possível utilizar técnicas de processamento de linguagem natural para classificar o risco de fraude de licitação de obras públicas com base nos textos de publicações da Seção 3 do Diário Oficial da União.

1.3 Objetivos

Nesta Seção são apresentadas as definições dos objetivos norteadores do trabalho, estes organizados e divididos em objetivo principal e secundário.

1.3.1 Objetivo Principal

O principal objetivo é encontrar uma definir 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.

1.3.2 Objetivos Secundários

O trabalho ainda possui os seguintes objetivos específicos:

1. Formação de um banco de dados de publicações da Seção 3 do Diário Oficial da União das últimas duas décadas.

2. Identificação (anotação do banco de dados) de publicações vinculadas a casos onde fora constatado risco à fazenda pública.

3. Melhorar o desempenho de classificadores ou construir um modelo capaz de servir de alerta de risco de fraude.

4. Encontrar forma eficiente de representação do texto do Diário Oficial da União.

1.4 Organização do Manuscrito

O trabalho foi estruturado da seguinte forma: O Capítulo 2 apresenta a Fundamentação Teórica que é a base deste trabalho e o Capítulo 3 descreve brevemente o projeto no qual está inserido esta pesquisa. O Capítulo 4 aborda os Trabalhos Relacionados recentes na determinação de conluio e na classificação de textos usando tecnologias processamento de linguagem natural (NLP). O Capítulo 5 apresenta a metodologia de desenvolvimento do trabalho. O Capítulo 6 apresenta os resultados alcançados pela pesquisa. Por último, o Capítulo 7 apresenta as conclusões derivadas dos resultados e as necessidades de trabalhos futuros para o projeto.

2

Fundamentação Teórica

"Não é sensato deixar um dragão fora dos teus cálculos se vives perto dele."
– J. R. R. Tolkien

Este capítulo aborda os conceitos básicos necessários para entender o que será classificado e a importância da classificação, além da base conceitual da classificação de texto em processamento de linguagem natural e a avaliação dos resultados. Além disso, será explicado o Projeto *Deep Vacuity* no qual esta pesquisa se inclui.

2.1 Convênios, Processo Licitatório, Contratos e suas Publicações

Para compreender a estrutura dos dados a serem classificados e a estrutura das publicações disponíveis nos diários oficiais é conveniente se iniciar pela estrutura de um processo licitatório.

O processo de realização de uma obra pública segue eventos em uma linha pré-estabelecida pela lei de licitações [17], lei do Regime Diferenciado de Contratações (RDC) [22] e lei de pregões [18] (Vide Seção 2.1.4). A Figura 2.1 mostra o fluxograma das etapas processuais desde a celebração do convênio até o encerramento do contrato.

Quando a contratação se dará por recursos próprios, ou seja, sem a realização de um convênio, o primeiro passo público do processo é a publicação do edital da licitação, caso o recurso seja proveniente de outro ente federado o convênio deve ser firmado e publicado previamente.

Cada processo licitatório pode gerar um ou mais contratos, usualmente em obras é feita uma licitação para cada contrato, mas não há obrigatoriedade legal. Um contrato pode sofrer alterações financeiras, de prazos ou de cláusulas contratuais, sendo assim celebrados e publicados "termos aditivos". Caso o contrato se encerre normalmente (com

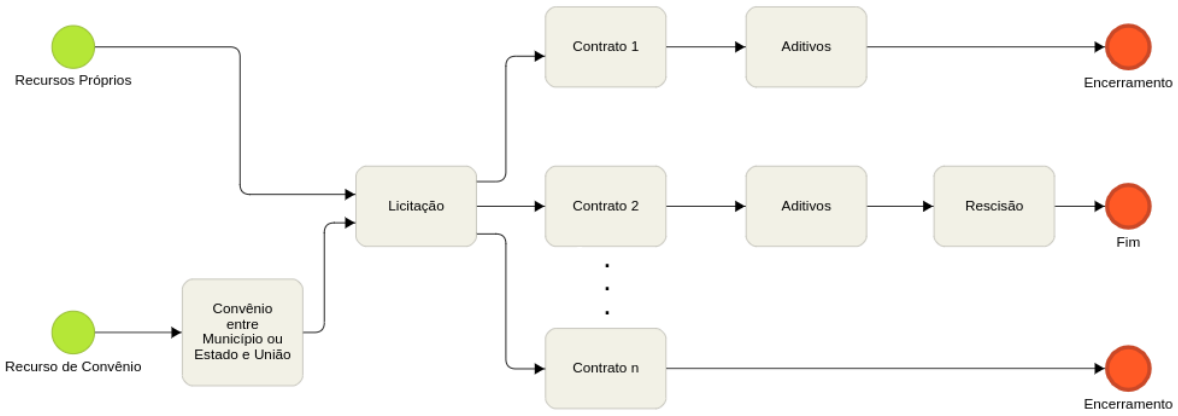


Figura 2.1: Fluxograma das etapas do convênio ao encerramento do contrato [17, 18]. Figura do Autor.

a execução do objeto) não há mais publicações, mas caso haja interrupção do contrato antes da sua conclusão é feita e publicada a rescisão contratual [17, 18].

2.1.1 Convênios

Os convênios ocorrem quando o recurso financeiro para realização da obra é de outro ente federado, ou seja, o recurso é passado da união para um estado ou município. Assim, é assinado um **convênio** para estabelecer as cláusulas de compromisso, prazos e valores. Usualmente um ministério ou uma autarquia é o órgão que promove o convênio (concedente) e um estado ou município o que recebe e executa o convênio (convenente).

Os convênios são regidos por uma série de normativos sendo o principal o Decreto 6.170, de 25 de julho de 2007 [21], que prevê em seu artigo 16:

> Art. 16. Os órgãos e entidades concedentes deverão publicar, até cento e vinte dias após a publicação deste Decreto, no Diário Oficial da União, a relação dos objetos de convênios que são passíveis de padronização.
>
> Parágrafo único. A relação mencionada no caput deverá ser revista e **republicada anualmente**.[grifado]

Neste documento se usa o termo "convênio" para denominar vários tipos de acordos ente entes federados, entre eles, de acordo com o definido por [26], os seguintes:

- "Transferência Voluntária (TV) - compreende a entrega de recursos correntes ou de capital a outro ente da Federação, a título de cooperação, auxílio ou assistência financeira. Nesse caso não se incluem aqueles decorrentes de mandamento constitucional, legal, os destinados ao sistema único de Saúde, bem como as descentralizações

de recursos a Estados, Distrito Federal e Municípios para a execução de ações cuja competência seja exclusiva da União."

- "Convênio – acordo, ajuste ou qualquer outro instrumento que discipline a transferência de recursos financeiros de dotações consignadas nos Orçamentos Fiscal e da Seguridade Social da União e tenha como partícipe, de um lado, órgão ou entidade da administração pública federal, direta ou indireta, e, de outro lado, órgão ou entidade da administração pública estadual, distrital ou municipal, direta ou indireta, ou ainda, entidades privadas sem fins lucrativos, visando a execução de programa de governo, envolvendo a realização de projeto, atividade, serviço, aquisição de bens ou evento de interesse recíproco, em regime de mútua cooperação;"

- "Contrato de Repasse - instrumento administrativo, de interesse recíproco, por meio do qual a transferência dos recursos financeiros se processa por intermédio de instituição ou agente financeiro público federal, que atua como mandatário da União;"

- "Termo de Execução Descentralizada - instrumento por meio do qual é ajustada a descentralização de crédito entre órgãos e/ou entidades integrantes dos Orçamentos Fiscal e da Seguridade Social da União, para execução de ações de interesse da unidade orçamentária descentralizadora e consecução do objeto previsto no programa de trabalho, respeitada fielmente a classificação funcional programática;"

- "Termo de Parceria - é o instrumento passível de ser firmado entre o Poder Público e as entidades qualificadas como Organizações da Sociedade Civil de Interesse Público destinado à formação de vínculo de cooperação entre as partes, para o fomento e a execução das atividades de interesse público, previstas no Art. 3o da Lei n 9.790, de 23 de março de 1999;"

- "Consórcio Público – São parcerias formadas exclusivamente por entes da Federação, na forma da Lei no 11.107, de 2005, para estabelecer relações de cooperação federativa, inclusive a realização de objetivos de interesse comum, constituída como associação pública, com personalidade jurídica de direito público e natureza autárquica, ou como pessoa jurídica de direito privado sem fins econômicos, regulamentado pelo Decreto nº 6.017/2007 [20]"

- "Transferência Legal - acordo, ajuste ou qualquer outro instrumento amparados por normativo legal, que discipline a transferência de recursos financeiros de dotações consignadas nos Orçamentos Fiscal e da Seguridade Social da União e tenha como partícipe, de um lado, órgão ou entidade da administração pública federal, direta ou indireta, e, de outro lado, qualquer pessoa física ou jurídica (pública ou privada com ou sem fins lucrativos), visando a execução de programa de governo, envolvendo a

realização de projeto, cuja relação jurídica não possa se constituir por um termo de convênio, um contrato de repasse, um termo de parceria, um acordo de cooperação técnica", um termo de compromisso ou um termo de execução descentralizada;"

Após a assinatura do convênio, ele é publicado na Seção 3 do Diário Oficial da União, assim como seus aditivos, um exemplo de publicação de convênio pode ser visto na Figura 2.2.

FUNDAÇÃO NACIONAL DE SAÚDE

EXTRATO DO TERMO DE COMPROMISSO Nº TC/PAC 0637/2014 E APROVAÇÃO FORMAL

Partícipes: Fundação Nacional de Saúde, CNPJ: 26.989.350/0001-16, situada no SAS, Quadra 4, Bloco N, 5° andar, Brasília/DF e o Município de Mãe d'Água/PB, CNPJ: 09.084.088/0001-41, situado na Rua Luiz Furtado de Figueiredo, 48, Mãe d'Água/PB. Objeto: Sistema de Esgotamento Sanitário. 1) Da Compromissária: R$ 1.422.143,41, sendo que sobre R$ 71.107,17 corendo a despesa à conta de dotação orçamentária consignada no Programa de Trabalho 10.512.2068.10GE.0001, UG 255000 Gestão 36211, conforme NE nº 2014NE000682 de 30/04/2014, Fonte 0151, ED 4440.42. Data de assinatura: 31/12/2014. Signatários: Antonio Henrique de Carvalho Pires, CPF: 767.810.894-04 e Margarida Maria Fragoso Soares, CPF: 041.626.334-87. Processo n° 25100.007.584/2014-51.

Figura 2.2: Publicação exemplificativa de um **extrato de convênio** (neste caso denominado termo de compromisso). Além do texto, ressalta-se a quantidade de informação em forma numérica: datas, CNPJ, CPF, valores monetários e outros identificadores. (Figura de [29])

2.1.2 Processo Licitatório

O processo licitatório é o processo público para a realização de obras públicas segue-se, na maioria das vezes, a Lei de Licitações 8.666/93 [17] e não a Lei dos Pregões 10.520/2002 [18] por esta última ser limitada a serviços comuns de engenharia [124] (Vide Seção 2.1.4).

A partir de 2011, as contratações de obras públicas também puderam ser realizadas pelo Regime Diferenciado de Contratações [22] em serviços que envolviam, principalmente, obras referentes a Copa do Mundo de Futebol de 2014 e Olimpíadas de 2016, sendo posteriormente alterada para a inclusão de outros tipos de projetos.

Na principal regulamentação, a lei de licitações [17], são citadas as modalidades de licitação:

Art. 22. São modalidades de licitação:
I - concorrência;

II - tomada de preços;
III - convite;

Destas modalidades, a concorrência é a que envolve maiores recursos e, por isso, tem controles com mais fases e, consequentemente, mais publicações.

São esperadas as seguintes publicações referente às licitações [124]:

- Aviso de licitação/edital
- Abertura da licitação
- Resultado do julgamento da habilitação
- Resultado de recursos ao julgamento da habilitação
- Abertura das propostas comerciais
- Resultado do julgamento das propostas comerciais
- Resultado de recursos ao julgamento das propostas comerciais
- Resultado Final

A Figura 2.3 mostra uma publicação no Diário Oficial da União de um resultado de julgamento final de uma licitação.

RESULTADO DE JULGAMENTO
CONCORRÊNCIA Nº 19/2001

O DNOCS - Departamento Nacional de Obras contraSecas,informa o resultado final - Concorrência
19/2001, conforme segue: consórcio vencedor formado pelas empresas: Andrade Gutierrez/EIT/OAS/BARBOSA MELLO.

JOSE FRANCISCO DOS SANTOS RUFINO
Diretor Geral

(SIDEC - 03/05/2002)

Figura 2.3: Exemplo de publicação no Diário Oficial da União de um resultado de julgamento (Final) de uma licitação [29]

2.1.3 Contrato

Findo o processo licitatório, inicia-se a contratação que deve ser publicada de acordo com o artigo 61 da Lei 8.666/93 [17].

> Art. 61. Todo contrato deve mencionar os nomes das partes e os de seus representantes, a finalidade, o ato que autorizou a sua lavratura, o número do processo da

> licitação, da dispensa ou da inexigibilidade, a sujeição dos contratantes às normas desta Lei e às cláusulas contratuais.
>
> Parágrafo único. A publicação resumida do instrumento de contrato ou de seus aditamentos **na imprensa oficial**, que é condição indispensável para sua eficácia, será providenciada pela Administração **até o quinto dia útil do mês seguinte** ao de sua assinatura, para ocorrer no prazo de vinte dias daquela data, qualquer que seja o seu valor, ainda que sem ônus, ressalvado o disposto no Art. 26 desta Lei. [grifado]

Além da publicação do contrato, ainda são esperadas as publicações de termos aditivos e, se for o caso, rescisão contratual. A Figura 2.4 mostra três exemplos de publicação no Diário Oficial da União de fases de um contrato, com o extrato do próprio contrato, um aditivo contratual e a rescisão contratual. Nesta figura, nota-se que apesar de o texto seguir uma padronização, ela não é rígida e estruturada, está sujeita a erros de digitação e outros erros como apresentação de uma empresa para a Contratação e quatro empresas em consórcio para a rescisão.

DEPARTAMENTO NACIONAL DE OBRAS CONTRA AS SECAS

EXTRATO DE CONTRATO Nº 9/2002

Nº Processo: 59400.004219/2001
Contratante: DEPARTAMENTO NACIONAL DE OBRAS CONTRA AS SECAS
CNPJ Contratado: 17262213000194
Contratado : CONSTRUTORA ANDRADE GUTIERREZ SA
Objeto: Execução das obras e serviços de cons trução da barragem Congonhas, tipo mista (CCR e Terra),incluindo fornecimento, instalação e mon tagem dos equipamentos hidromecanicos e eletri cos, localizado no Municipio de Grão Mogol, no Estado de Minas Gerais.
Fundamento Legal: Lei nº 8.666/93
Vigência: 07/06/2002 a 24/05/2005
Valor Total: R$ 249.722.329,29

Fonte de Recurso	Nota de Empenho
100000000	2002NE900465

Data de Assinatura: 05/06/2002

(SICON - 06/06/2002) 193002-11203-2002NE000508

EXTRATO DE TERMO ADITIVO Nº 9/2009

Número do Contrato: 9/2002. Nº Processo: 59400005308200847. Contratante: DEPARTAMENTO NACIONAL DE OBRAS CONTRA AS SECAS. CNPJ Contratado: 17262213000194. Contratado : CONSTRUTORA ANDRADE GUTIERREZ SA -Objeto: Adequar a Planilha Contratual às determinações exaradas nos Acórdãos nº 1.774/2004-TCU/PLENÀRIO e nº 2.110/2006-TCU/PLENÁRIO e 1.803/2008-TCU/PLENÁRIO. Fundamento Legal: Lei n] 8.666/93 Valor Total: R$85.685.516,10. Data de Assinatura: 19/02/2009.

(SICON - 20/02/2009) 193002-11203-2009NE900047

EXTRATO DE RESCISÃO Nº 1/2011

Espécie: Rescisão de Contrato PGE-09/2002; Partes: Departamento Nacional de Obras Contra as Secas, CGC nº 00.043.711/0001-43 e o Consórcio Andrade Gutierrez/Oas/Barbosa Mello/EIT; Objeto: Rescisão do Contrato PGE-09/2002, cujo objeto consiste na execução de obras e serviços de construção da Barragem Congonhas, incluindo o fornecimento, instalação e montagem dos equipamentos hidromecânicos e elétricos; Data da Assinatura: 28/12/2011; Assina: Elias Fernandes Neto, Diretor Geral do DNOCS; Fundamento Legal: Lei nº 8.666/93, alterada pela Lei nº 8.883/94.

Figura 2.4: Exemplo de publicação no Diário Oficial da União de: (a) Contratação da Construtora Andrade Gutierrez S.A. (b) Aditivo para adequação das cláusulas contratuais e (c) Rescisão que informa Consórcio formado por três empresas ao invés de empresa única (Figuras de [29]).

2.1.4 Nova Lei de Licitações

Em 1º de abril de 2021, for sancionada a Lei de Licitações e Contratos Administrativos (nova lei de licitações) sob o número 14.133/2021 [25] em substituição às leis 8.666/93 [17] (Licitações), 10.520/2002 [18] (Pregões) e a lei 12.462/2001 [22] (Regime Diferenciado de Contratações). Estas últimas leis ainda valerão por dois anos.

Espera-se que haja mudanças na forma de publicação dos documentos referentes a licitações públicas, mas esta regulamentação e o sistema previsto nesta lei (Painel Nacional de Contratações Públicas) ainda não foram lançados.

2.2 Processamento de Linguagem Natural

Uma das áreas da inteligência artificial é o processamento de linguagem natural (NLP, sigla em inglês para *Natural Language Processing*) no qual são desenvolvidas técnicas e modelos que usam textos produzidos por humanos como entrada ou como saída [70]. Muitas tarefas podem ser desenvolvidas neste campo dentre as quais se encontram ferramentas usadas no nosso dia-a-dia como tradutores automatizados, *chatbots*, filtros de *spam* ou corretores gramaticais.

De acordo com Goldberg [70] as principais áreas de NLP atualmente são:

- **Análise de Sentimentos** (*Sentiment Analysis*): usada para interpretar e classificar o tipo de emoção tem um texto, frase ou documento.
- **Modelos de Linguagem** (*Language Modeling*): predição da próxima palavra ou letra de um texto. Modelos atuais são capazes de gerar textos complexos a partir de poucas palavras de modo que o resultado final é quase imperceptivelmente artificial.
- **Tradução** (*Machine Translation*): aplicações de tradução automática como o Google Translate[1].
- **Classificação de Textos** (*Text Classification*): Veja Seção 2.3 a seguir.
- **Respostas a Perguntas** (*Question Answering*): normalmente com base em um texto de referência um modelo é treinado para responder perguntas.
- Outros: Reconhecimento de entidades nomeada, sumarização, extração de relacionamento etc.

[1] https://translate.google.com.br/

2.3 Classificação de Textos

A Classificação de Textos é a tarefa do ramo do campo de NLP onde se elabora modelos para a atribuição de classes ou categorias a sequências textuais ou documentos.

Segundo Bramer [15], a classificação, sendo uma atividade frequente, na maioria das vezes envolve dividir elementos em categorias mutualmente exaustiva e excludentes, ou seja, cada objeto só tem uma categoria ou classe.

Muitos processos decisórios podem ser formulados como problemas de classificação[15], ou seja, a ação a ser tomada sobre um elemento depende exclusivamente da classe na qual este é classificado.

A classificação de textos tem aplicações que se estendem da detecção de *spam*, determinação de fonte ou autor, etc. As classificações podem ser binárias, de duas classes, ou de múltiplas classes.

Para cada técnica, a definição dos termos documento, palavra e frequência apresentado será utilizado o mesmo termo da técnica inicial, de forma a preservar o trabalho, mesmo que seja sinônimo com outro elemento do texto. Assim, documento será o elemento a ser classificado, no caso o extrato de publicação do Diário Oficial.

As palavras são definidas como os *tokens* que formam a vetorização e a frequência é a contagem relativa das palavras.

A classificação de textos por uma função linear é definida [70]:

$$f(x) = x \cdot W + b \tag{2.1}$$

Onde x é o vetor de entrada e a matriz W e o vetor b são os parâmetros. O objetivo do aprendizado é determinar W e b para que a entrada $x_{1,k}$ (k é o tamanho do vetor de entrada) produza a saída desejada, ou seja, as n classes no vetor $y_{1,n}$, dado pela saída da função $f(x)$.

Para a classificação de risco desta pesquisa, tendo uma classificação binária, a saída da função $f(x)$ é um escalar que pertence a $[0; 1]$, sendo os valores 0 e 1 atribuído a cada uma das classes.

Para alcançar a classificação há um *pipeline* que começa pelo **pré-processamento**, passando pela **extração de características**, pela **modelagem**, pela **avaliação** até ser obtido o modelo de **classificação** de texto [81]. Este processo pode ser ilustrado pela Figura **??**, onde é destacada a necessidade de reformulação dos procedimentos de pré-processamento, extração de características e modelagem, face a uma avaliação não satisfatória. Os conceitos básicos destas etapas serão abordados a seguir.

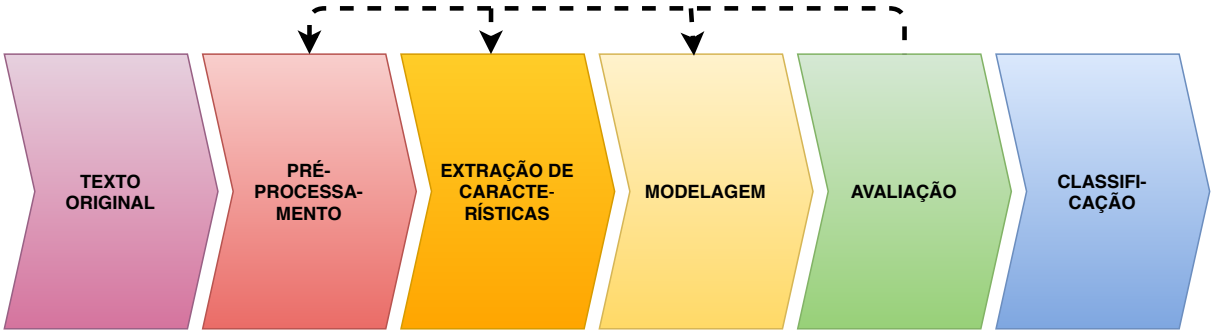


Figura 2.5: *Pipeline* do processo de classificação de textos. Após a avaliação dos resultados pode ser necessário reavaliar todos os processos anteriores. (Figura do Autor, adaptada de Kowsari et al. [81])

2.3.1 Pré-processamento de texto

A etapa de pré-processamento do texto visa balancear a eliminação de elementos e informações do texto que não contribuem para o modelo, a redução da complexidade e a manutenção das informações relevantes [56].

A extração de características por não se tratar da classificação em si, pode ser considerada dentro da etapa de pré-processamento [36], mas aqui é tratado à parte na Seção 2.3.2.

Os primeiros passos que foram estudadas e aplicadas no pré-processamento foram [127]:

- Remoção de pontuação
- Remoção de acentos e caracteres especiais
- Remoção de formatação HTML
- Remoção de números
- Minusculização
- Remoção de *stop-words*
- *Stemming* e Lematização
- *N-gram*

Alguns dos passos não são indicados para todos os tipos de processamento de texto, por exemplo a minusculização dificulta tarefas de extração de entidades nomeadas, mas no caso de classificação estes processos podem ser utilizados de acordo com Denny and Spirling [56], principalmente ao se utilizar características baseadas em cestos de palavras (Vide Seção 2.3.2).

Remoção de Pontuação e Números

A remoção de pontuação elimina os caracteres que não são letras. Se não forem eliminados os números, deve-se atentar para a alteração de sentido dos numerais ao se eliminar o indicador de parte fracionária (ex.: vírgula decimal em português ou o ponto em inglês) [13].

Remoção de Acentos e Caracteres Especiais

A remoção de acentos e caracteres especiais, ou seja, a transformação de caracteres acentuados como "â" em "a" ou "ç" em "c" que ajuda na eliminação de erros de acentuação e a remoção de símbolos como "º" ou "&" [13]. Entretanto este processo pode trazer um ruído de duplicidade de significados de parônimos[2], como, por exemplo:

- Para (verbo parar) e Pará (estado)
- Da (contração de+a) e dá (verbo dar)
- Fábrica (local) e fabrica (verbo fabricar)

Remoção de Formatação HTML

A formatação HTML presente em vários extratos de textos e documentos obtidos de repositórios da Internet e por *crawlers* pode ser removida com facilidade, uma vez que a formatação HTML é disposta entre os símbolos "<" e ">", é utilizada para a inclusão de links, imagens, formatação de página, formatação de tabelas etc., dados que normalmente trazem pouca ou nenhuma informação para o contexto de classificação de textos. Esta apresentação do formato facilita sua remoção uma vez que estes símbolos são raramente utilizados em textos fora do contexto de equações matemáticas [13].

Remoção de números

A remoção de números pode reduzir consideravelmente a dimensionalidade de vetores de banco de palavras no contexto de textos oficiais, uma vez que cada documento, processo, lei, página etc. possui um número determinado, além disso, os documentos são datados e são apresentados muitos valores financeiros [13]. Cada um destes valores numéricos incrementam a dimensionalidade da vetorização.

Uma alternativa a perda destas informações será apresentada na Seção 5.5 na página 67.

[2] "É a relação que se estabelece entre palavras que possuem significados diferentes, mas são muito parecidas na pronúncia e na escrita". https://www.soportugues.com.br/secoes/seman/seman7.php, visitado em 02 de setembro de 2020.

Minusculização

O processo de minusculização visa evitar a criação de vetores separados para a mesma palavra, por esta ter sido apresentada em caixa diferente [13].

Perde-se com isso a capitalização de nomes próprios e junta-se os vetores destes com substantivos comuns, como acontece com o nome Coelho e o animal ou com a cidade de Palmas e o aplauso.

Remoção de *Stop-words*

Várias palavras utilizadas trazem pouca informação direta ao contexto de uma classificação e tem uma frequência de ocorrência muito grande [77], ou seja, sua presença não traz informação sobre as diferentes classes, estas são palavras comuns ou *stop-words*, termo cunhado por Luhn [91].

Em português, costumam fazer parte da lista de *stop-words* os artigos "o, a, as, os, um, uma, uns, ...", as conjunções "e, que, quais, mas, cujo, ...", as conjugações de verbos muito usados como "ter", "estar" e "haver" entre outras.

Stemming e Lematização

O processo de *stemming* é a redução das palavras apenas ao seu tronco (*stem*), assim palavras como "derivando" ou "derivar" são reduzidas a uma forma comum "deriv", este processo pode ser alcançado removendo os sufixos verbais e adverbiais mais comuns da língua (*Suffix stripping)*, entretando este método pode causar muitos erros quando o final de uma palavra coincide com um sufixo [77].

A lematização pode ser definida no contexto lexicográfico[3] como [45]:

> "[...] modo de agrupamento padrão das diversas variantes de um mesmo signo, com a finalidade de simplificar a apresentação e desse modo facilitar a consulta dos extratos lexicais em geral. Nos dicionários práticos, a lematização consiste em encontrar um item, isto é, uma forma gráfica representativa de todas as formas que uma unidade de significação lexicográfica (tradicionalmente palavra ou palavras compostas) pode tomar. É assim que o infinitivo é geralmente escolhido para simbolizar todas as formas do paradigma verbal (ex.: Ter por tenho, temos, terei, etc.); o masculino singular representa todas as formas do paradigma nominal e do paradigma adjetivo".

Assim, a lematização, tendo como objetivo definir o *lema* de cada palavra pode seguir um processo computacional muito mais complexo ao estabelecer para cada uma das

[3] "[„,] ramo da linguística que se ocupa do estudo do vocabulário de uma língua, visando essencialmente a forma e a significação das palavras para a elaboração de dicionários, léxicos e terminologias". https://www.infopedia.pt/dicionarios/lingua-portuguesa/lexicografia, visitado em 07 de setembro de 2020.

palavras seu contexto, sua classificação gramatical em uma tarefa de NLP denominada *Part-of-speech (POS) tagging*, ou seja, anotação de classes de palavras em um corpus [101].

N-gram

Quando se utiliza o conjunto de palavras (*bag-of-words*) ao invés da sequência textual em uma tarefa de processamento de linguagem natural perde-se as informações referentes a relação de cada uma das palavras com as palavras que a cercam. Uma forma de contrabalançar esta deficiência da técnica é utilizar como *token*, além da palavra isolada, as ocorrências de conjunto de palavras em pares, *bigrams*, ou trios, *trigrams* e assim por diante [32].

Há de ser ter em conta que a utilização de *n-grams* cresce exponencialmente com n, segundo a equação:

$$n(t - 2(n-1)) + \sum_{i=1}^{n-1} 2i \quad n, t \in \mathcal{N} \tag{2.2}$$

Onde t é o número de elementos na base de dados.

Por conta disso, tende-se a trabalhar com o menor dimenção n de *grams* possível.

2.3.2 Extração de Características

A extração de características é a transformação de informações de um domínio específico para outro, como forma de possibilitar a modelagem. O domínio de origem pode ser um sensor, uma imagem, uma tabela de dados ou **um documento de texto** [70].

A decisão das formas de extração de características é fundamental para o sucesso do aprendizada de máquina [70]. O texto normalmente apresenta informações de maneira desestruturada e com vocabulário bastante amplo, assim uma das dimensões da vetorização tem dimensão na ordem de dezenas a centenas de milhar (tamanho do vocabulário).

Bag of Words

Um atributo comum para uma classificação de textos é o *Bag of Words* (BOW) ou Cesto de Palavras que é um conjunto que lista todas as palavras (*tokens*), sem repetição, que ocorrem em um texto, com a possibilidade de se trabalhar *n-grams*, normalmente sendo chamado neste case de *bag of n-words* [70].

Como afirmado antes, neste "cesto" as palavras são guardadas sem que seja armazenada a estrutura do texto ou a interconexão entre as palavras dentro dele. A representação desta característica pode ser feita pela presença booleana de um *token*, em forma de histograma ou com vetores *one-hot*.

TF-IDF

Uma forma mais elaborada de colocar o peso em cada uma das palavras é considerar a frequência relativa das palavras em relação aos documentos pelo método conhecido como TF-IDF (***T**erm **F**requency - **I**nverse **D**ocument **F**requency*) [93].

O procedimento inicia ao representar a palavra w em um documento d apenas por sua frequência normalizada em cada documento:

$$\frac{\#_d(w)}{\sum_{w' \in d} \#_d(w')} \tag{2.3}$$

O TF-IDF altera esta frequência multiplicando pelo inverso do número de documentos em que esta palavra ou *n-gram* ocorre:

$$\frac{\#_d(w)}{\sum_{w' \in d} \#_d(w')} \times \log \frac{|D|}{|\{d \in D : w \in d\}|} \tag{2.4}$$

A utilização deste peso para o BOW ressalta termos que são distintivos de um determinado documento e termos que possuem alta frequência terão pesos menores [3].

Representações Vetorias de Palavras - Word2VEC

Uma outra forma de transformar o texto em uma informação numérica vetorial é utilizar um modelo de rede neural (não profunda) para treinar a vetorização de palavras em um determinado *corpus* [94]. Isso gera um modelo da língua conhecido como *word embedding*.

Uma das principais limitações desta representação [37] é que cada palavra só recebe um vetor, assim apenas o sentido predominante da palavra é representado.

2.3.3 Modelos de Classificação Lineares

O princípio da classificação supervisionada é fazer com que a máquina aprenda com exemplo e seja capaz de generalizar o aprendizado. O problema de classificação linear de textos pode ser representado, de acordo com Manning and Schutze [92], na forma $(\vec{x}, c)$, onde $\vec{x} \in \mathbb{R}^n$ é o vetor de medidas e c é a classe. Para o problema binário um modelo de classificador linear seria:

$$g(\vec{x}) = \vec{w} \cdot \vec{x} + w_0 \tag{2.5}$$

Onde seria atribuída uma classe c_1 para $g(\vec{x}) > 0$ e uma classe c_2 para $g(\vec{x}) \leqslant 0$. Sendo os parâmetros deste modelo o vetor $\vec{w}$ e o limiar w_0.

Já no modelo de classificação binária a Equação 2.5 é simplificada uma vez que a dimensão de saída é unitária, assim, w se torna um vetor e b uma constante:

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + w_0 \tag{2.6}$$

Pode-se aplicar a $f(\vec{x})$ uma função *sinal* para que o domínio de saída deixe de estar $]-\infty, +\infty[$ para ficar com apenas duas saídas: -1 e $+1$.

A Equação 2.6 mostra x que é proveniente da vetorização obtida na extração de características (Seção 2.3.2) de dimensão d_n podendo representá-la como:

$$f(x) = \sum_{i=0}^{d_n} (x_i \cdot w_i) + b \tag{2.7}$$

Para se obter uma saída da função que represente a probabilidade da classe ao invés da sua predição costuma-se [70] aplicar uma função sigmoide:

$$\sigma(f(x)) = \frac{1}{1 + e^{-(x \cdot w + b)}} \tag{2.8}$$

A classificação linear pode ser obtida por outros métodos e abordagens, como Naïve Bayes, *suport vector machines* (SVM) etc., entretanto não será possível abordar todos aqui apesar de no Capítulo 6 haver a aplicação de vários deles.

2.3.4 Redes Neurais

Nas últimas duas décadas as redes neurais têm se mostrado ferramentas com grande capacidade e potencial, entretanto no domínio de classificação de textos os métodos clássicos ainda se mostram competitivos (como em [52] por exemplo).

A base de desenvolvimento das redes neurais foram e, em parte, são os perceptrons. *Perceptron* é um classificador binário definido por:

$$f(x) = \begin{cases} 1 & \text{se } w \cdot x + b \geqslant 0 \\ 0 & \text{caso contrário} \end{cases} \tag{2.9}$$

O perceptron por ser um classificador linear se limita a problemas linearmente separáveis (o que exclui qualquer problema com estrutura não linear), assim foram desenvolvidas os perceptrons multicamadas onde cada um dos neurônios k é definido por uma função:

$$f_k(x) = \sum_i (w_k i \cdot x + b_k i) \tag{2.10}$$

Onde w_n são os pesos aplicada a cada uma das entradas e b_n é o viés, vide Figura 2.6.

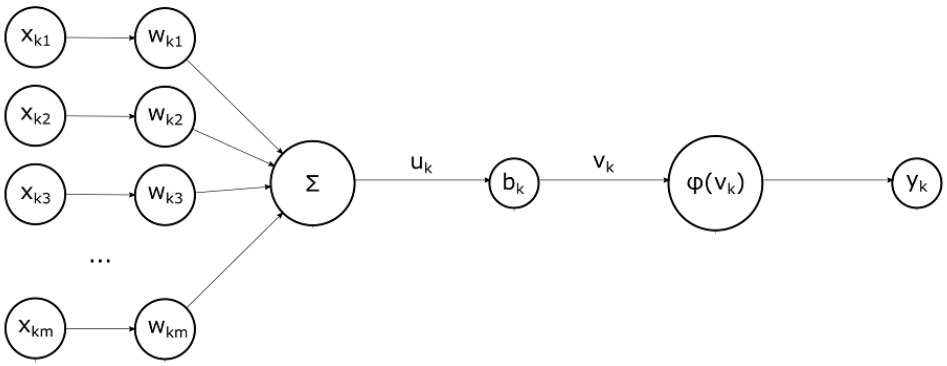


Figura 2.6: Esquema de entrada e saída de dados de um neurônio em uma rede neural artificial. [Figura do autor, adaptada de [70]]

Função de Ativação

A saída da função $f_k(x)$ está nos limites $]-\infty, +\infty[$ e para limitar as saídas em um determinado intervalo e reduzir a linearidade da solução e é aplicada a função de ativação, sendo as principais [34]:

Sigmoide limita a saída ao intervalo $[0, 1]$, sendo que os valores muito negativos tendem a 0 e os valores muito positivos a 1. Seu uso tem se reduzido atualmente pois se o gradiente local for pequeno há tendência de perda de sinal e a função não é centrada no zero. É definida pela função:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.11}$$

Tangente hiperbólica (tanh). É centrada no zero, uma vez que tem $[-1, 1]$ como intervalo de saída e pode ser vista como uma função derivada da sigmoide:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma(2x) - 1 \tag{2.12}$$

ReLU (*Rectified Linear Unit*) é uma função que aumenta a convergência em alguns casos e de cálculo bem simples por se resumir a $f(x) = \max(0, x)$, ou seja, a função toma valor zero se a entrada é negativa e toma o valor da entrada em caso contrário, assim seu intervalo de saída é $[0, \infty[$. Há uma variação da ReLU chamada de *Leaky* ReLU (Relu com vazamento) onde há um sinal de saída de valores negativos reduzido por um fator α que visa combater a tendência de algumas redes terem neurônios mortos (com valores presos em zero). A função *Leaky* ReLU é dada por:

$$f(x) = \begin{cases} \alpha x & \text{se } x < 0 \\ x & \text{se } x \geqslant 0 \end{cases} \tag{2.13}$$

Normalmente a camada de saída de uma rede neural recebe como função de ativação a função **Softmax** [129]. Esta função recebe um conjunto de K valores reais e os transforma num conjunto de K valores que somam 1. Deste modo a saída da função permite que os valores sejam interpretados, apesar da falta de calibração, como probabilidade e é dada pela equação:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.14}$$

Onde $\vec{z}$ é o vetor de entrada.

Função de Perda

O objetivo do treinamento é aproximar o valor predito $\hat{y}$ do valor de referência y, com isso chega-se a função de perda (*loss function*) representada por $L(\hat{y}, y)$. Como a perda é calculada para cada uma das amostras de treinamento, calcula-se a perda média $\mathcal{L}$ para um conjunto de parâmetros Θ:

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i, \Theta), y_i) \tag{2.15}$$

Quando se busca o valor mínimo de $\mathcal{L}$, se obtém o objetivo do treinamento. Goldberg [70] complementa que, para evitar o *overfiting* adiciona-se a $\mathcal{L}$ na Equação 2.15 um termo de regularização $R(\Theta)$ com peso λ;

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i, \Theta), y_i) + \lambda R(\Theta) \tag{2.16}$$

Várias funções de perda (ou funções de custo, dada por L) podem ser aplicadas entre elas pode-se destacar **Hinge** ou perda SVM definida, com $\tilde{y} = sinal(\hat{y})$, como $L(\tilde{y}, \hat{y}) = max(0, 1 - \tilde{y} \cdot \hat{y})$ e ***Binary cross entropy*** definido tendo duas classes: 0 e 1 e a saída da função de classificação passando por uma função sigmoide gerando a função:

$$L(\tilde{y}, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \tag{2.17}$$

Para a regularização, o valor de λ é definido manualmente caso a caso e o valor de R é dependente dos pesos do vetor $\vec{w}$ da Equação 2.5. A formas comuns de cálculo do R são norma-L1 (RL_1), norma-L2 (RL_2)e *elastic-net* (R_{EN}):

$$RL_1(\vec{w}) = \|\vec{w}\|_1 = \sum_i |w_i| \tag{2.18}$$

$$RL_2(\vec{w}) = \|\vec{w}\|_2^2 = \sum_i |w_i|^2 \tag{2.19}$$

$$RL_{EN}(\vec{w}) = \lambda_1 RL_1(\vec{w}) + \lambda_2 RL_2(\vec{w}) \tag{2.20}$$

Uma forma eficiente de se resolver a otimização de mínimo da Equação 2.16 é por meio de SGD (*Stochastic Gradient Descent*) onde para uma determinada amostra são computados os gradientes para correção dos parâmetros Θ e a função de perda é tradada como uma função dos parâmetros Θ, sendo eles alterados em uma taxa de aprendizado [70].

Backpropagation

Na Equação 2.16 foi mostrado a necessidade de se computar os parâmetros do modelo Θ, no caso dos perceptrons multicamada eles são w_k e b_k da Equação 2.10. Para este procedimento é necessário que o erro calculado volte para o início da rede corrigindo os parâmetros (normalmente iniciados aleatoriamente), a isso é dada a denominação de *backpropagation* [71].

Tendo a função de ativação dada por g, a saída da camada n dada por $h^{(n)}$, cada camada será calculada por:

$$h^{(n)} = g^{(n)}(w^{(n)}h^{(n-1)} + b^{(n)}) \tag{2.21}$$

Sendo que a entrada da primeira camada ($h^{(0)}$ em $n = 1$) são os dados de entrada x e a saída da última ($n = N$) camada é $\hat{y}$.

Com o objetivo de reduzir a função de custo $\mathcal{L}(y, \hat{y})$, os pesos devem ser ajustados de acordo com sua derivada em relação à função custo:

$$\frac{\partial \mathcal{L}}{\partial W_{jk}^{(i)}} \tag{2.22}$$

Onde $W_{jk}^{(i)}$ representa o peso da rede aplicado ao valor que vai do nó (ou neurônio) j na camada $i-1$ para o nó k na camada i.

Para se calcular o ajuste nos pesos é necessário aplicar a regra da cadeia[4]. A derivada da função de custo (tendo a saída limitada ao intervalo $[0, 1]$) é:

$$\frac{d\mathcal{L}}{d\hat{y}} = \hat{y} - y \tag{2.23}$$

Também deve-se aplicar a derivada da função de ativação, sendo ela a sigmoide $S(x) = 1/(1 + e^{-x})$, esta derivada é dada por:

$$\frac{dS}{dx} = S(x)(1 - S(x)) \tag{2.24}$$

Assim, pode-se aplicar a regra da cadeia a estas derivadas junto com a Equação 2.21:

$$\frac{\partial \mathcal{L}}{\partial W^{(n)}} = \frac{d\mathcal{L}}{d\hat{y}} \frac{\partial \hat{y}}{\partial h^{(n+1)}} \frac{\partial h^{(n+1)}}{\partial W^{(n)}} \tag{2.25}$$

Se essas correções fossem executadas a cada amostra treinada o modelo dificilmente encontraria o mínimo e nem teria capacidade de generalização. Assim, os ajustes feitos são os da média de um *batch* ou de um conjunto de treinamento.

Taxa de Aprendizado

Os pesos calculados na *backpropagation* são, ainda, ajustados por uma taxa de aprendizado (*learning rate*) a cada *batch* de treinamento. A taxa de aprendizado permite que haja um ajuste na forma como se busca o mínimo da função de custo.

Como a determinação da taxa de aprendizado tende a ser empírica, uma forma de otimizar o processo de treinamento é utilizando taxas cíclicas de aprendizado, como exposto por Smith [118].

Em seu trabalho no ano seguinte, Smith [119] sugere que o ciclo que aprendizado seja único, ou seja, que durante todas as épocas só haja uma subida e uma descida na variação da taxa de aprendizado que ficou conhecido como *One Cycle Policy*, como exposto na Figura 2.7.

Dropout

Em redes neurais, utiliza-se como regularização o *dropout* [120] que é uma forma simples de se evitar o *overfitting*. Neste caso, são avaliados os nós da rede que pouco contribuem para os dados de saída (*output*) e estes nós são eliminados do modelo final.

[4]Regra de cálculo que permite calcular a derivada de uma função usando a derivada de sua subfunções, é apresentada como: $\frac{dx}{dy} = \frac{dx}{dz}\frac{dz}{dy}$

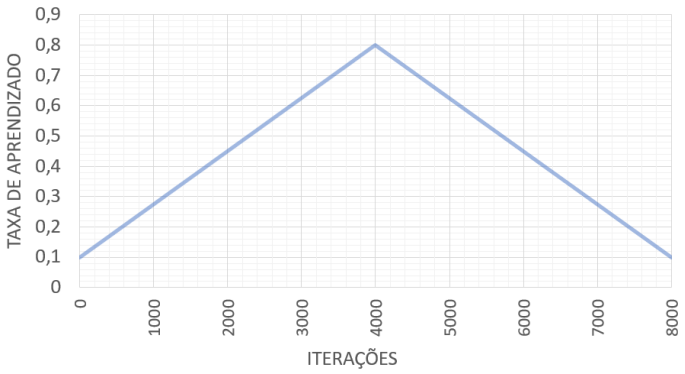


Figura 2.7: Ilustração da *One Cycle Policy* de Smith [119]. [Figura do autor, adaptada de [119]]

2.3.5 Tipos de Rede Neurais

Existem uma infinidade de variações e classificações de redes neurais, aqui serão abordadas as principais redes que foram testadas neste trabalho: redes *autoencoders*, redes recorrentes e *Long Short-Term Memory (LSTM)*.

Redes *Autoencoders*

Redes neurais do tipo *Autoencoder* [82] são redes neurais que possuem na sua arquitetura um gargalo - por isso também são chamadas de *bottleneck* -, forçando a rede a desenvolver nesta camada um descritor mínimo das características necessárias à execução da tarefa, no caso a classificação de textos.

A variação mais moderna deste tipo de rede, a *Variational Autoencoder* [79] insere na camada do gargalo uma camada probabilística com média e desvio padrão, possibilitando que seja criado um modelo gerador, ou seja que é capaz de gerar dados similares aos dados de entrada.

Redes Neurais Recorrentes (RNN)

Redes neurais recorrentes são redes que trabalham com uma dimensão a mais, geralmente chamada de dimensão temporal, no caso de textos ressalta-se as arquiteturas LSTM (*Long-Short-Term Memory*) que serão aprofundadas mais a frente neste trabalho.

As RNN permitem que modelos linguísticos prevejam a "próxima palavra" com base no histórico de toda a frase, possibilitando a criação de geradores de texto.

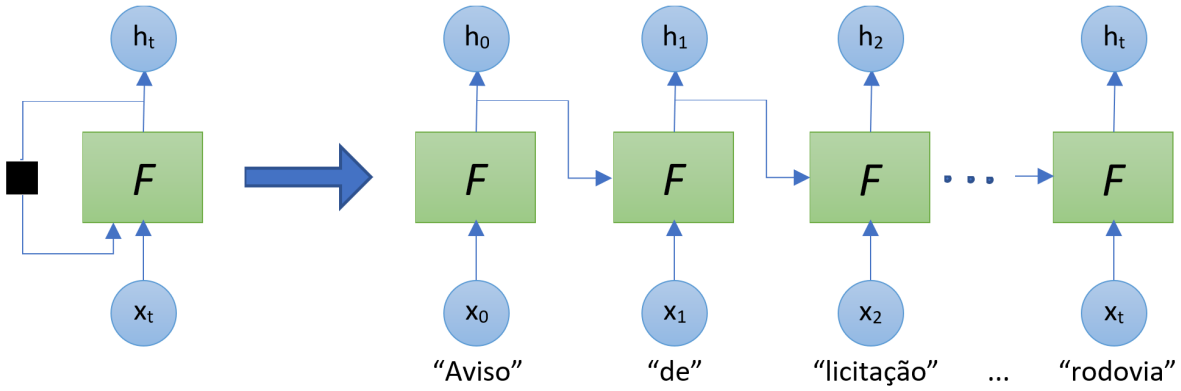


Figura 2.8: Representação das relações temporais (entre palavras) em uma rede neural recorrente. [figura do autor]

Se é dada uma sequência ordenada de vetores $x_{1:n} = x_1, x_2, \ldots, x_n$ a saída da função recorrente RNN na posição i sendo $1 \leqslant i \leqslant n$[5] assumirá o valor recorrente de:

$$\hat{y}_i = h_i = RNN(\hat{y}_{i-1}, x_i) \tag{2.26}$$

Long Short-Term Memory (LSTM)

No desenvolvimento das redes recorrentes durante a década de 90, existia o desafio de lidar com a “explosão” ou com o “desaparecimento” dos sinais que eram enviados de volta no tempo (*backpropagation in time*) fazendo com que os pesos ficassem oscilando, levando tempo demais atingir o aprendizado ou sem que chegasse a este ponto. Assim, para contrapor a estes desafios, foi proposta a arquitetura LSTM em 1997 por Hochreiter and Schmidhuber [75], a primeira rede recorrente a apresentar mecanismos de portões (*gating*) que são funções matemáticas que simulam portões lógicos [70].

A Figura 2.9 representa a diferença entre uma RNN e uma rede LSTM. A rede neural recorrente promove uma junção do sinal de saída do elemento anterior (Função $A(\hat{y}_{i-1}, x_i)$ da Figura 2.9b), já a rede LSTM apresenta uma série de interações entre estes valores para permitir uma otimização da “memória” dos sinais.

A principal ideia das LSTM é a utilização de um “corredor” que permite a passagem do sinal na dimensão temporal com a utilização de *gates* (x e $+$ na linha horizontal da

[5] Goldberg [70] afirma que por questão de simplicidade o vetor inicial s_0 é assumido como um vetor de zeros.

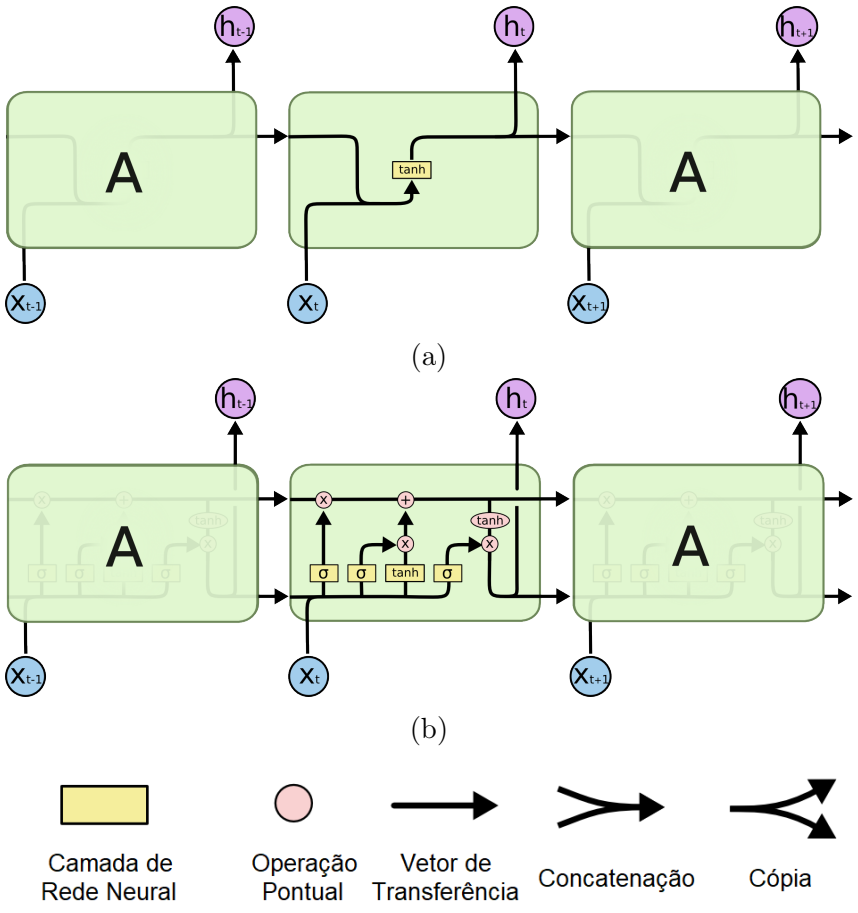


Figura 2.9: Comparação entre uma RNN tradicional (a) e uma rede LSTM (b) com a legenda dos operadores (c). [Figura de Olah [98]]

Figura 2.9-b) [98].

$$
\begin{aligned}
s_j = R_{LSTM}(s_{j-1}, x_j) &= [c_j, h_j] \\
c_j &= f \odot c_{j-1} + i \odot z \\
h_j &= o \odot \tanh(c_j) \\
i &= \sigma(x_j W^{xi} + h_{j-1} W^{hi}) \\
f &= \sigma(x_j W^{xf} + h_{j-1} W^{hf}) \\
o &= \sigma(x_o W^{xo} + h_{j-1} W^{ho}) \\
z &= \tanh(x_z W^{xz} + h_{j-1} W^{hz})
\end{aligned}
\tag{2.27}
$$

A Equação 2.27 define matematicamente a arquitetura LSTM no início do tempo j. Sendo os vetores c_j a componente de memória que faz parte do "corredor" e h_j que é a saída oculta (*hidden*). Os portões são definidos por i (entrada), f (esquecimento) e o (saída).

Uma característica importante desta arquitetura é o portão de esquecimento que é dado por f que varia de 0 a 1 por ser tratar de uma função sigmoide que no cálculo de c_j a entrada do "corredor" (c_{j-1}) é multiplicado por f.

A implantação de uma rede recorrente normalmente se dá no sentido positivo do tempo, ou seja, na sequência de leitura do texto, entretanto as relações textuais não são unidirecionais e a interpretação do sentido que se dá a um termo pode depender de um contexto apresentado após ele. Desta forma, o modelo pode percorrer os dados no sentido inverso ou, ainda, em ambos os sentidos paralelamente.

Assim, foi proposto por Graves and Schmidhuber [73] um modelo classificatório baseado em LSTM bidirecional (BiLSTM obtendo relutados mas satisfatórios que RNNs e LSTMs em dados de reconhecimento de fala. As saídas dos canais de ida e volta (*foward and backward*) podem ser unidas por operações matemáticas (soma, multiplicação, diferença, ...) ou serem concatenados em um vetor maior.

2.3.6 Rede Neural Multicanal

As redes neurais multicanais (*Multistream Neural Networks*) vêm sendo utilizadas em muitas aplicações de inteligência artificial, uma vez que permitem a utilização de processamento e aprendizado paralelo de vários tipos de modelos [1, 49, 54, 64, 106].

Nestas redes a camada de entrada é dividida em duas ou mais redes que depois são fundidas em um único canal que leva aos dados de saída. Esta fusão dos canais (*fusion*

stage) pode se dar por, virtualmente, qualquer operador matemático, mas usualmente usa-se:

- Soma
- Subtração
- Multiplicação
- Média Aritmética
- Quadrado da Diferença
- Módulo da Diferença
- Concatenação

As possibilidades de implantação de redes multicanal são inúmeras uma vez que a separação pode se dar em qualquer lugar da rede, pode acontecer várias vezes e pode-se mesclar as formas de fusão dos canais.

2.3.7 Conjuntos de Treinamento, Validação e Testes

Segundo Goldberg [70], quando se lida com problemas de aprendizado de máquina, deve-se observar um conjunto de entrada $x_{1:k}$ e os rótulos correspondentes $y_{1:k}$. Ao produzir uma função $f(x)$ que mapeia x para $\hat{y}$, deve haver uma forma de avaliar as métricas que comparam y com $\hat{y}$.

O objetivo de elaboração da função f é a capacidade de generalização, ou seja, a capacidade de prever corretamente $\hat{y}'$ quando comparado a y' dado um conjunto desconhecido x'.

Alguns modelos permitem que esta separação siga a solução *leave one out* (deixar um de fora), onde o modelo é treinado apenas para o conjunto $x_{1:j-1,j+1:k}$ gerando uma função f_j, válida para avaliar x_j e chegando a $\hat{y}_j$ que será comparado a y_j.

Modelos que não permitem ou por sua estrutura ou pela demanda computacional utilizar as técnicas de *leave one out* [70], podem ser treinados separando o conjunto x e y em um conjunto para treinamento e outro para teste de forma randômica. Normalmente, estes conjuntos são separados na proporção de 80%/20% até 70%/30%. Estes números são derivados do princípio de Paretto [100].

Quando, ainda, há a necessidade de acompanhamento da qualidade do treinamento durante as escolhas de parâmetros, como no caso de redes neurais, utiliza-se a separação nos conjuntos de treinamento, *validação* e testes, sendo o tamanho dos conjunto de validação e treinamento da ordem de 20% dos itens, a ilustração deste processo está na Figura .

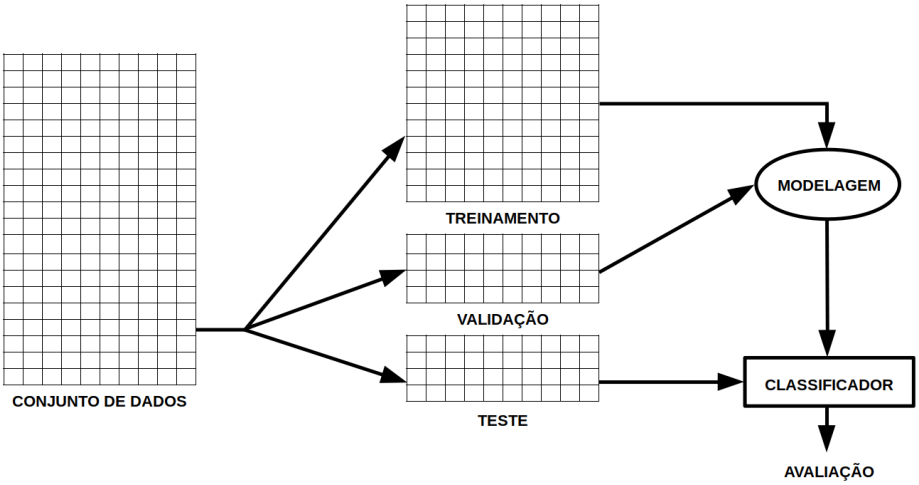


Figura 2.10: Divisão dos elementos em conjuntos de treinamento, validação e testes. [Figura do autor, baseada em Bramer [15]]

Validação Cruzada

Quando se realiza a técnica de *leave one out* se está comparando todas as possibilidades de treinamento com predição, ou seja, são observados todos cenários de agrupamento do conjunto de treinamento, entretanto quando são utilizados os conjuntos de treinamento (com ou sem validação) e teste, normalmente é inviável treinar todas as combinações possíveis de agrupamento destes conjuntos.

Desta forma, para que o se proceda a diminuição do erro derivado desta limitação pode-se repetir a divisão, de forma aleatória, dos conjuntos num processo de validação cruzada k vezes (*k-fold cross-validation*). A quantidade de repetições normalmente se encontra na ordem de grandeza de uma dezena.

Segundo Bramer [15], sendo s a estimativa de acurácia de uma predição para uma predição e N o número de elementos no conjunto de testes, o erro padrão é dado por $\sqrt{s(1-s)}/N$.

Quando se realiza a validação cruzadas k vezes o erro padrão é reduzido para:

$$\frac{\sqrt{s(1-s)}}{kN} \tag{2.28}$$

2.4 Avaliação dos resultados

Para medir a performance de um classificador binário têm-se vários parâmetros que podem ser utilizados, mas primeiro se tem que separar os resultados pela aderência que eles tem aos dados reais (*groundtruth)*[93]. Para isso são definidos quatro tipos possíveis de resultados para uma amostra classificada no conjunto de testes: Verdadeiro Negativo

(TN), Verdadeiro Positivo (TP), Falso Positivo (FP), Falso Negativo (FN), contabilizando a ocorrência destes tipos cria-se a matriz de confusão de acordo com a Tabela 2.1. Normalmente a matriz de confusão é apresentada em percentual de amostras.

Neste caso considera-se como "positivo" as amostras que apresentam Risco = 1.

Tabela 2.1: Tabela de avaliação de um classificador binário. Esta avaliação gera a quantidade de cada um dos campos, ou seja, a matriz de confusão, normalmente apresentada após normalização.

	Risco 0 foi predito	**Risco 1 foi predito**
Risco 0 é correto	Verdadeiro Negativo	Falso Positivo
Risco 1 é correto	Falso Negativo	Verdadeiro Positivo

Com o quantitativo pode-se calcular a precisão (*precision*), a sensibilidade (*recall*) e a acurácia (*accuracy*):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.29}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{TN}} \tag{2.30}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{TP} + \text{TN}}{\text{Total}} \tag{2.31}$$

Com a média harmônica da precisão e da sensibilidade calcula-se o *F1-Score*:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.32}$$

Estes três critérios são acompanhados para todos os modelos e testes.

Durante o treinamento das redes neurais, utiliza-se frequentemente a medida de AUC (*Area Under Curve*) do gráfico ROC (*Receiver Operating Characteristic*).

Como a classificação binária é feita na escala $[0, 1]$, o **limiar** (*threshold*) é aquele que abaixo do qual a amostra é classificada como negativa e acima do qual é classificada como positiva.

Para cada limiar l nos limites do classificador $[0; 1]$, obtém-se valores diferentes de Falsos Positivos e Verdadeiros Positivos gerando a ROC. A representação da curva ROC num gráfico como o da Figura 2.11 onde os eixos são a Taxa de Falsos Positivos e e Taxa de Verdadeiros Positivos.

Como apresentado na Figura 2.11, um classificador aleatório gera uma curva ROC linear entre os pontos $0; 0$ e $1; 1$ e com área sob esta curva (AUC) de $0, 5$. Por outro lado, um classificador perfeito classificaria todas as amostras corretamente no intervalo de limiar ℓ sendo $\{\ell \in \Re | 0 < \ell \leqslant 1\}$ e, assim, chega-se a uma área sob a curva de $1, 0$.

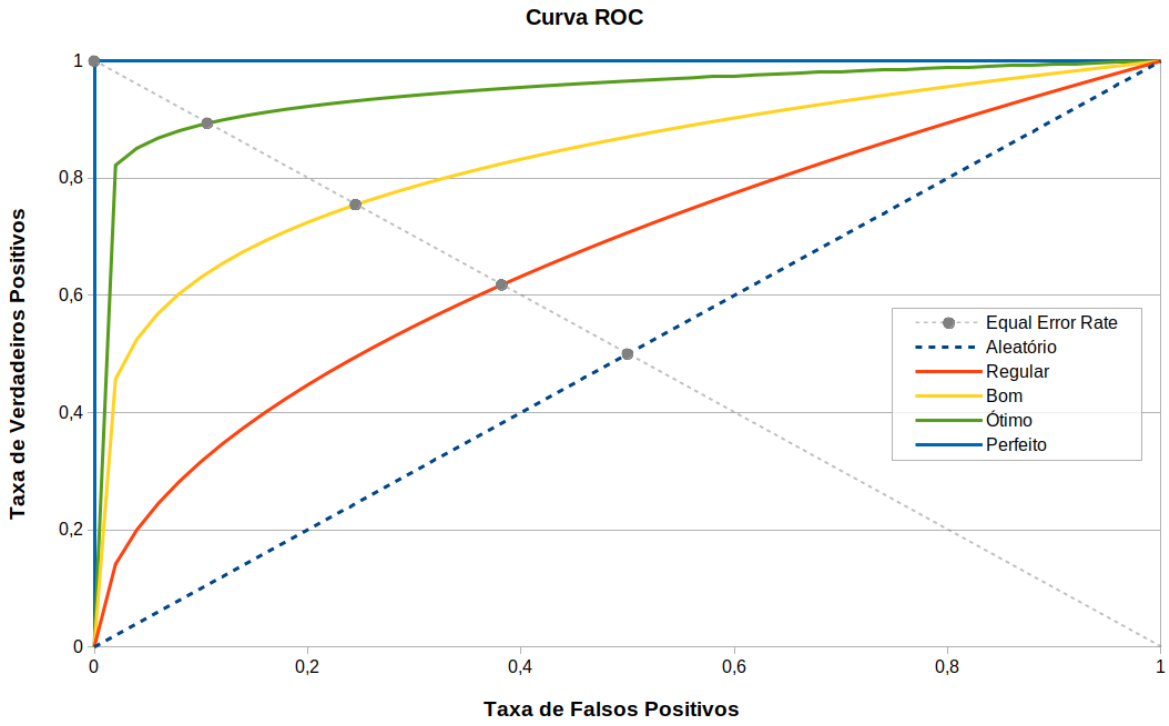


Figura 2.11: Representação da curva ROC (*Receiver Operating Characteristic*) com diferentes limiares de classificação. Também é mostrado a indicação do cálculo do *Equal Error Rate* conforme Seção 2.5 [Figura do autor adaptada de Draelos [59]]

Desta forma, avalia-se o classificador durante o treinamento pelo critério de AUC que varia de $0,5$ a $1,0$.

2.5 *Equal Error Rate* - EER

Após a saída do preditor, ou seja, do valor de probabilidade do classificador no intervalo $[0; 1]$, é necessário definir um limiar decisório (*threshold*) para definir uma publicação como de risco, ou não.

Este limiar é definido como o valor no qual a taxa de rejeição é igual a taxa de aceite. A taxa de rejeição é a proporção de entradas válidas que são rejeitadas, ou seja, é a taxa de falsos negativos e a taxa de aceite é a proporção de entradas incorretamente aceitas, ou seja, são os falsos positivos [43].

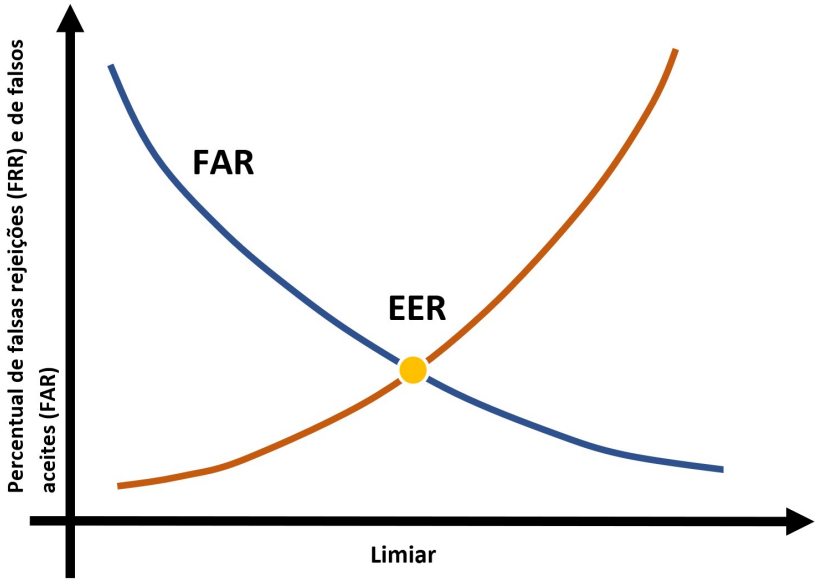


Figura 2.12: Gráfico para cálculo do EER - *Equal Error Rate*, onde FRR é a taxa de rejeição e FAR é a taxa de aceite. [Figura de [43]]

O cálculo da EER é feito pela computação da taxa de falsos positivos e negativos que são obtidos para cada limiar. Este processo é visualizado na Figura 2.12. Um procedimento similar pode ser executado obtendo o valor que da curva ROC que cruza a reta $f(x) = x - 1$ na Figura 2.11 [43].

Este capítulo abordou os conhecimentos básicos que são utilizados como base para o desenvolvimento deste trabalho. O capítulo seguinte apresenta as principais características do *framework* computacional do projeto *Deep Vacuity*.

3

Projeto *Deep Vacuity*

"My soul unraveled out of mental
The shell returns to dust
I focus on the present concentrate on what I find."

– Gojira

Neste Capítulo são apresentadas as principais características do *framework* computacional do projeto *Deep Vacuity*[1], este sendo o motivador e o inspirador das atividades desenvolvidas e apresentadas neste manuscrito, sendo elas parte integrante do projeto.

O projeto *Deep-Vacuity* tem como objetivo principal o desenvolvimento de metodologias para identificação de comportamento de cartéis de empresas em obras públicas utilizando técnicas de aprendizado de máquina e inteligência artificial.

Dentre as atividades relacionadas ao projeto, tem-se como atividades a serem desenvolvidas:

- Desenvolvimento e implantação de um *framework* com infraestrutura computacional de alto desempenho distribuído, para dar suporte às atividades de monitoramento e detecção de atividades suspeitas em licitações públicas. Estas realizadas em bases públicas disponíveis em sítios da web;
- Propor metodologias de captação de dados públicos nas esferas federais, estaduais e municipais;
- Incorporar no *framework* desenvolvido a base de aprendizado e *expertise* do corpo pericial da Polícia Federal no processo de detecção de fraudes em obras públicas para os ambientes computacionais;

[1]Este nome foi escolhido pela combinação do termo *Deep*, devido ao tipo de análise profunda das informações processadas e extraídas, estas provenientes de bases públicas ainda não exploradas; e o termo *Vacuity*, seguindo a descrição da letra da música homônima da banda francesa *Gojira* do albúm *The way of all flesh*. Maiores informações sobre o significado do termo *Vacuity* podem ser encontradas neste link: https://www.letras.mus.br/gojira/1343928/traducao.html.

- Permitir ações que abranjam outras esferas do poder público, tanto regulatório, quanto de fiscalização, incluindo forças policiais e órgãos de controle.

O projeto *Deep-Vacuity* visa auxiliar tarefas de fiscalização, auditoria e investigação, estas atualmente realizadas em sua maioria por análise humana. Desta forma, é de grande utilidade que o sistema tenha um canal de interação com um tipo de usuário final. Espera-se então que o sistema completo tenha a capacidade acoplar interfaces úteis e de *frontend* com fácil usabilidade, como seguem alguns exemplos:

- Visualização de dados, metadados, agrupamentos e seleção de entidades;
- Painel de controle com gráficos e informações em escala macro e micro de determinados campos de conhecimento;
- Visualização de grafos e correlações entre dados, documentos e entidades;
- Representação espacial bidimensional ou tridimensional das relações entre dados;
- Visualização dos mesmos dados em diferentes escalas e tipos de intervalos temporais
- Representações dos dados por sua geo-localização;
- Interação com o usuário para validação e refinamento de resultados dos modelos inteligentes;
- Capacidade de retroalimentação de dados no sistema para retreinamento de modelos existentes ou treinamento de novos modelos inteligentes.

Este desenvolvimento é fruto da parceria com Instituto Nacional de Criminalística (INC), do Serviço de Perícia em Engenharia (SEPENG) do Departamento de Polícia Federal (PF). A seguir serão listadas as principais informações das atividades que estão sendo realizadas no escopo do projeto.

3.1 Histórico e Motivação

Em meados de 2014, surge a maior operação de combate à corrupção da história da república brasileira, denominada de **Operação Lava-jato** [42, 83], a qual atingiu uma magnitude extraordinária em valores financeiros movimentados, e também na quantidade de agentes públicos envolvidos, de diversas naturezas e espectros políticos.

Com o advento desta operação, observou-se notoriamente que a capacidade de investigação, completamente limitada e manual, mitigou o desenvolvimento e o surgimento de diversas linhas de trabalhos científicos e e de iniciativas populares, para o auxílio no combate aos crimes de corrupção (também conhecidos como "crime do colarinho branco").

Estas modalidades estão agora mais sofisticados e complexas (vide Departamento de Operações Estruturadas da Empresa Odebrecht [66]), se comparado aos tradicionalmente relacionados na literatura sobre o tema[2], como comparativo da evolução e melhoria do *modus operandi* dos envolvidos.

Mediante essa avalanche de novos dados a serem processados, bem como pelo motivo nobre da causa em combater, surgiram vários movimentos na área de tecnologia da informação com o objetivo explícito de auxiliar no combate à corrupção, a saber:

1. **Operação Serenata de Amor**

   A proposta da Operação Serenata de Amor [96] é atuar no monitoramento dos gastos referentes à atividade parlamentar, principalmente monitorando de forma automatizada, com auxílio de tecnologia, os reembolsos efetuados pela Cota para Exercício da Atividade Parlamentar (CEAP) – verba que custeia alimentação, transporte, hospedagem e até despesas com cultura e assinaturas de TV dos parlamentares.

2. **Operação Política Supervisionada**

   A Operação Política Supervisionada [12] fiscaliza de forma detalhada os gastos realizados via CEAP (ou CEAPS). Até o momento já foram economizados mais de R$ 5,5 milhões do dinheiro público graças a estas fiscalizações e às exigências feitas diretamente aos parlamentares para que devolvam o dinheiro público indevidamente utilizado.

   A OPS conta com a ajuda de seus colaboradores, espalhados pelo Brasil, para o levantamento de informações necessárias para a conclusão de fiscalizações, como por exemplo, o envio de fotos de endereços suspeitos em diversas cidades do país. Além disso, qualquer um pode ser um fiscal dos gastos públicos e este site oferece dados suficientes para isso.

3. **Observatório da Despesa Pública**

   O Observatório da Despesa Pública (ODP) [47] é uma unidade permanente do Ministério da Transparência e Controladoria-Geral da União (CGU) voltada à aplicação de metodologia científica, apoiada em tecnologia da informação de ponta, para a produção de informações que visam a subsidiar e a acelerar a tomada de decisões estratégicas por meio do monitoramento dos gastos públicos.

   O objetivo do ODP é contribuir para o aprimoramento do controle interno e funcionar como ferramenta de apoio à gestão pública, os resultados gerados pela unidade servem como insumo para realização de auditorias e fiscalizações conduzidas pela

[2] Vide exemplo da Ação Penal 470 - conhecida como "Mensalão" [99]

CGU, bem como para informar aos gestores sobre indicadores gerenciais relativos à realização de gastos públicos, de modo a permitir análises comparativas, subsidiando a tomada de decisões para melhoria da aplicação dos recursos públicos.

3.2 Desenvolvimento do Projeto do Sistema *Deep Vacuity*

Paralelamente ao desenvolvimento desta pesquisa, é desenvolvido o sistema que fará gestão do banco de dados, aplicação dos modelos desenvolvidos e possibilitará a interface a usuário destes dados. Para isso o projeto emprega metodologia de gerenciamento de projetos baseadas nos princípios preconizados por métodos ágeis e SCRUM[3], com adoção de ferramenta própria para suporte das atividades de gerenciamento.

A Figura 3.1 apresenta a proposta de arquitetura inicial do *Deep Vacuity*, onde são apresentados os principais aspectos da parte de infraestrutura e sua organização para o funcionamento esperado. No caso, o funcionamento é voltado para dar suporte à utilização de técnicas de aprendizagem de máquina profundo na atividade pericial.

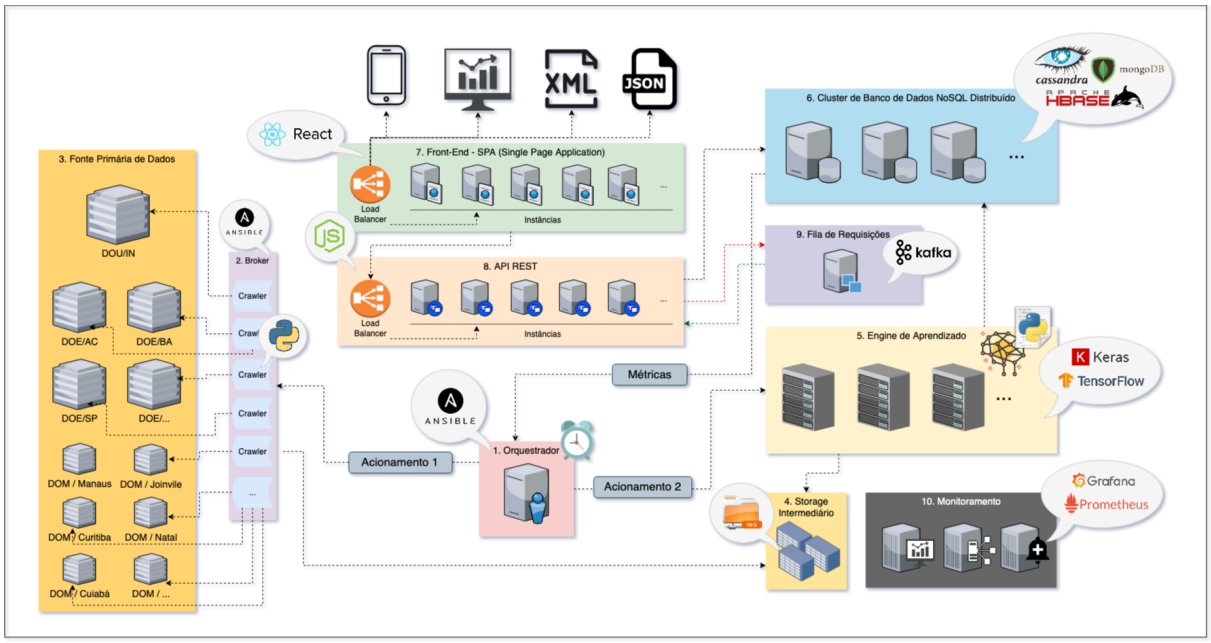


Figura 3.1: Arquitetura inicial proposta para o Sistema *Deep Vacuity*. Esta figura foi elaborada por Leonardo Carvalho como parte do Relatório Descritivo de Atividades e Produtos Desenvolvidos (documento não publicado).

A aplicação de metodologia de gestão de projetos baseada na visão do SCRUM deve criar um equilíbrio entre as demandas de escopo, tempo, custo, qualidade e bom relaci-

[3]SCRUM: https://scrum.org.

onamento entre os diversos atores do projeto. O sucesso dessa gestão estará relacionado ao alcance dos objetivos de: entrega dentro do prazo previsto, dentro do custo orçado, com nível de desempenho adequado, com plena aceitação pelo cliente e seus representantes (usuários finais), com atendimento de forma controlada às mudanças de escopo e em respeito à cultura da organização.

Por meio de um trabalho coordenado e interdependente entre as equipes do Instituto Nacional de Criminalística da polícia Federal e da Universidade de Brasília, as etapas de cada fase serão planejadas, discutidas, executadas e documentadas. As tarefas e atividades do convênio são sempre supervisionadas pelos coordenadores das duas instituições.

As equipes operacionais responsáveis pelos esforços técnicos serão lideradas por pesquisador sênior e por um gerente operacional. Uma equipe de gestão e qualidade será responsável pelos processos de gestão, pelo aceite dos relatórios do projeto e pelo acompanhamento do projeto. As equipes operacionais serão formadas por profissionais com diferentes experiências e qualificações.

Este capítulo abordou as características computacionais do projeto *Deep Vacuity*. O capítulo seguinte apresenta os trabalhos atuais relacionados a esta pesquisa.

4

Trabalhos Relacionados

"Todo aquele que ouve as minhas palavras e as põe em prática
é como o homem prudente que edificou a sua casa sobre a rocha."
– Evangelho segundo São Mateus

Neste capítulo serão abordados trabalhos de destaque dos últimos anos no enfrentamento do problema de conluio em licitações públicas e nas técnicas e ferramentas aplicáveis à formação e classificação dos dados.

Desta forma, haverá duas seções: a primeira tratará de abordagens específicas na tentativa de se criar um mecanismo de detecção de conluio na Seção 4.1, e, em seguida serão passadas as técnicas mais recentes de inteligência artificial que dão base à formulação das soluções buscadas neste projeto, na Seção 4.2.

4.1 Detecção de Conluio em Licitações

Nesta seção serão abordadas as iniciativas e técnicas atuais utilizadas na detecção de fraudes e conluio em licitações, especialmente aquelas realizadas pela Polícia Federal e a CGU, assim como a utilização de técnicas de inteligência artificial para este fim.

4.1.1 Iniciativas da Polícia Federal na Detecção de Conluio

A Polícia Federal tem envidado esforços no desenvolvimento de mecanismos de detecção e comprovação de conluio em licitações. Neste intuito, Vallim [125] desenvolveu um modelo baseado em casos (CBR - *Case-based reasoning*) abordando as licitações de pavimentação. Estes serviços, por serem grandes demandantes de recursos públicos em todas as esferas de governo, são constantemente alvo de de ações criminais. O modelo baseou-se em uma estrutura de dados que considerava: o tipo de licitação, as empresas

envolvidas, os contratos e dados georeferenciados (localização) focando na classificação de casos de conluio e levantamento manual das informações.

Em outra abordagem, vários estudos conduzidos por peritos da Polícia Federal se basearam no comportamento do conjunto de competidores que atuaram na manipulação das propostas utilizando ferramentas de estatística para identificação e comprovação do conluio em licitações.

Apesar da capacidade comprobatória destes métodos econométricos, eles são limitados a um único mercado ou órgão/empresa pública dominante e a um período limitado de tempo, tendo sido utilizados em com sucesso em contratos da operação Lava Jato [113, 114] e projetos de infraestrutura [115].

4.1.2 Iniciativas da Controladoria Geral da União na Detecção de Conluio

Várias iniciativas da Controladoria Geral da União, que é o órgão de controle interno do Governo Federal, também desenvolveram pesquisas no sentido de alcançar um classificador confiável para a detecção de fraudes em licitação.

Ralha and Silva [105] desenvolveram técnica de mineração de dados multi-agente (MAS) e descoberta de conhecimento em banco de dados (*database knowledge discovery*) com o intuito de detectar a cartelização de mercados públicos se baseando no sistema de compras governamentais ComprasNet[1].

Foi proposta solução utilizando mineração com regra de associação (ARM) sob a base de compras do ComprasNet do período de 2005 a 2008, num total de 26.615 entradas, sendo 2.701 licitações. Segundo os autores, houve uma correta identificação de formação de cartel em 90% dos casos.

Balaniuk et al. [9] focou na avaliação de risco de fraude em agências governamentais usando classificadores do tipo Naïve Bayes para o planejamento de auditoria. Foram utilizados dados estruturados e busca por padrões de atividade fraudulenta. O processo separou 2.560 pares de entidades públicas e privadas dentro de um universo de 795.954 pares com alta propabilidade de risco.

Sun and Sales [122] utilizaram redes neurais tradicionais e redes neurais profundas (*Deep Neural Networks* - DNN) para elaborar um sistema de alarme antecipado. Estes estudos usaram, de maneira geral, como características (*features*) e indicativos de fraude: o número de propostas, a relação entre preço e custo estimado, a relação entre os agentes públicos e privados, vinculação a partidos políticos etc.

[1] Acessível em https://www.gov.br/compras/pt-br/, visitado em novembro de 2020

Carvalho and Carvalho [38] encontraram bons resultados com o uso de modelos bayesianos com dados estruturados de penalidades aplicadas por órgãos públicos. Foram utilizados dados dos servidores públicos, seus cargos públicos, salário, número de contas julgadas irregulares, número de contas julgadas regulares nos órgãos públicos e servidores relacionados.

4.1.3 Estimativa do risco de contratos na Paraíba

Uma forma diferente de estimar o risco de contratos foi abordada por de Menezes [53] e se baseia na identificação de contratos rescindidos de uma determinada empresa e associação desta a um risco de contratação, com os contratos podendo ser separados por área e incluindo **obras e serviços de engenharia**.

Foi adotara técnica de *data augmentation* baseada na técnica de *oversampling* SMOTE[39] (*Synthetic Minority Oversampling Technique*) e utilizados apenas modelos clássicos lineares.

Foram alcançados índices de F1-*score* de 74,5% ao se utilizar o modelo *tree bag*.

4.1.4 Caso de Contratos de Rodovias na Polônia

Anysz et al. [7] selecionaram apenas 249 registros de concorrências que incluía como dados: nome dos participantes, local, valores das propostas, forma de execução, escopo e tipo de rodovia.

Foi utilizado com indicativo de quatro critérios de conluio da OCDE:

- Número de concorrentes que apresentaram propostas;
- Diferença de valor entre as propostas;
- Repetição da ordem de contratação para o mesmo contratado no mesmo local; e
- Mesmo conjunto de propostas em poucas licitações.

Para cada um dos critérios foram determinados limites numéricos para definir se seria caracterizado um indício de conluio. Uma crítica que pode ser feita ao método é que são apenas levantadas características externas para a elaboração dos casos o que leva apenas a uma descrição superficial de cada item do conjunto de dados e não há uma indicação mais robusta da existência de conluio.

Assim, foram aplicados dois métodos de classificação: um sistema *neuro-fuzzy* e classificação de atributos por redes neurais.

O método por *neuro-fuzzy* teve uma taxa de acerto de 64,28% no conjunto de teste/-validação. Entretanto quando se faz a estratificação por grupos a taxa de acerto para "conluio muito esperado" foi de apenas 25%.

Para as redes neurais foi utilizado o *Statistica 13* com uma *hidden layer* tendo de 3 a 30 neurônios. Mesmo utilizando uma rede muito simples fora obtido 90% de acerto no conjunto de validação/teste. Este trabalho trouxe uma forma tradicional de se encarar este tipo de problema com boas ideias para se quantificar o conluio quando não se tem acesso a dados mais detalhados dos procedimentos licitatórios, entretanto o conjunto de dados foi pequeno e as técnicas de classificação limitadas e pouco descritas.

4.1.5 Uso de NLP para avaliação de interferência privada na elaboração de regulamentos no Canadá

O artigo *Public Interest in the Regulation of Competition, Evidence from Wholesale Internet Access Consultations in Canada* de Reza Rajabiun e Catherine Middleton [104] foi publicado no Journal of Information Policy em 2015 e trata da análise textual em comunicações de empresas privadas ao órgão regulador canadense visando aferir a existência de aspectos que **reduziriam a competitividade**.

O trabalho analisou o conteúdo e os metadados de 25.797 cartas utilizando no processamento de linguagem natural as ferramentas do *software* Leximancer.

O artigo perpassa por resultados com grafos obtidos pelas técnicas e conclusões que podem ser obtidas pelo processamento.

Conclui que as técnicas de processamento de linguagem natural utilizadas são úteis, mas não forneceram resultado adequado sem supervisão humana. Mostrou o potencial de encontrar padrão de conluio na análise de linguagem natural com um banco de dados de textos de tamanho maior do que seria avaliável por uma equipe considerável de humanos em um tempo inferior e obtendo padrões difíceis de serem detectados sem o uso de inteligência artificial.

No trabalho a ser realizado, tendo em vista a sua utilização em processos criminais, ainda será necessário um desenvolvimento e conhecimento das técnicas de inteligência artificial (principalmente por redes neurais) e NLP para que um processo com nenhuma ou pouca supervisão humana possa prosperar.

4.1.6 Métodos Semânticos para Reutilizar LOD de Aviso de Licitação Públicas Europeias

O artigo *Semantic Methods for Reusing Linking Open Data of the European Public Procurement Notices* elaborado por Alvarez et al. [5], percorre a utilização de métodos semânticos de dados de licitações públicas como forma de modelar dados não-estruturados, enriquecer o sistema de classificação de produtos, publicar informações relevantes extraídas pela abordagem e utilizar algoritmos avançados para explorar as informações geradas.

O estudo utilizou dados do *Tenders Electronic Daily* que seria equivalente eletrônico europeu do Diário Oficial da União. As bases apresentavam 5 desafios: informação dispersa, duplicidade de avisos, inconsistência de formatos, ambiente multilíngue e dados de licitações de baixo valor. Destes desafios apenas a variação linguística não é observada no Brasil.

A pesquisa propôs uma arquitetura que permitiu que os dados fossem compilados e organizados em um *endpoint* SPARQL.

Uma boa porção da pesquisa com base nas licitações públicas brasileiras a ser implementada segue por caminhos semelhantes de *data mining* e encontra as mesmas dificuldades em se trabalhar com bases e textos não-estruturados.

4.1.7 Análise dos Métodos Utilizados da Detecção de Conluio

Vários métodos de mineração de dados podem ser escolhidos para a formação e organização do banco de dados. Foi observado que em países mais desenvolvidos os dados públicos já são apresentados de forma mais bem estruturados e que o sistema em desenvolvimento deve estar preparado para se adaptar a um modelo de *Linking Open Data*.

Um desafio comum entre vários trabalhos foi a consistência dos dados obtidos e a forma de interpretar de forma não-supervisionado ou semi-supervisionada os erros de digitação, duplicidade etc. que compõem a variabilidade inerente a uma base de dados de licitações públicas.

Por não possuir uma forma de diferenciar, dentro da base de dados, os casos com ou sem risco de conluio, a detecção de risco estive limitada nos trabalhos a comportamentos suspeitos.

O Processamento de Linguagem Natural não é frequentemente usado para classificar documentos de licitação quanto a risco ou fraude. A tecnologia é usada para avaliar o risco de fraude em sinistros de saúde [102, 126] e relatórios financeiros [69, 111]. Os dados das publicações de obras públicas não são uniformes o suficiente para serem estruturados e, mesmo se possível, seria extremamente trabalhoso e poderia ser feito ao custo de perder algumas características desconhecidas ou não detectadas.

Todos esses estudos fazem a classificação de risco em licitações baseada principalmente em dados estruturados e o uso da Processamento de linguagem natural para esse tipo específico de classificação é raro, superficial ou inexistente.

4.2 Classificação de Textos

Nesta seção serão abordados os trabalhos que tratam dos métodos, modelos e técnicas atuais utilizadas na tarefa de classificação de textos, conforme abordado na Seção 2.3.

4.2.1 Extração de Características - Descritores

Os métodos abordados na subseção 2.3.2 tem aplicação em estudos atuais, Kowsari et al. [81] identificou um amplo uso de TF-IDF como recurso para classificação de texto e, como arquitetura, o uso de redes neurais profundas, convolucionais e de crença profunda e BiLSTM.

No trabalho de Deng et al. [55] foram levantadas as técnicas atuais de extração de características: modelos de filtragem (Frequência de Documentos, TF-IDF, Ganho de Informação etc.), modelos *wrapper* (só foi observada funcionalidade com pequeno número de características), modelos de *embedding* (não foi encontrado modelo dedicado a classificação de textos) e modelos híbridos. Conclui que, apesar do grande número de técnicas de extração de características textuais existente, poucos são dedicados à classificação de textos e que os modelos de filtragem são os mais eficientes e foram os mais estudados para a categorização de textos.

4.2.2 Modelos de Classificação de Textos

A relação em trabalhos gerada pelo aplicativo *Connected Papers*[2] exposta na Figura 4.1, contém uma amostra das principais pesquisas desenvolvidas no campo de classificação de textos, a qual orientará o texto nas próximas Seções.

Vários trabalhos presentes na Figura 4.1 são de revisões sistemáticas, mas os trabalhos desenvolvidos por Sun et al. [121] e Liu et al. [88] envolvem a utilização do modelo de linguagem BERT e são abordados na Seção 4.2.3. Já os modelos bidirecionais desenvolvidos por Liu and Guo [87] (BiLSTM) e desenvolvido por Shen et al. [112] são expostos a seguir, entre outros.

Chen et al. [40] usou um modelo DNN com um extração de características 2D TF-IDF para classificar os comentários do Twitter sobre *cyberbullying* e discurso de ódio, e Chen et al. [41] classificou as avaliações dos clientes usando uma rede BiLSTM seguida por uma CNN 1D com características de *word emdeddings*, quando foi usado conhecimento desenvolvido por Kiperwasser and Goldberg [80] com a aplicação de redes BiLSTM.

Braz et al. [28] propôs, no já conhecido projeto Victor do Supremo Tribunal Federal [62], um modelo baseado em Bi-LSTM para classificar os documentos da suprema corte brasileira em seis classes. Foi verificado pelos autores que bastava alimentar a rede BiLSTM com os primeiros 1.000 *tokens* dos documentos, pois não havia degradação dos resultados quando a classificação era baseada nas duas primeiras páginas dos documentos (em PDF).

[2]Elaborado pelo autor utilizando o endereço web-curto: `shorturl.at/pABCI`, construído em fevereiro de 2021.

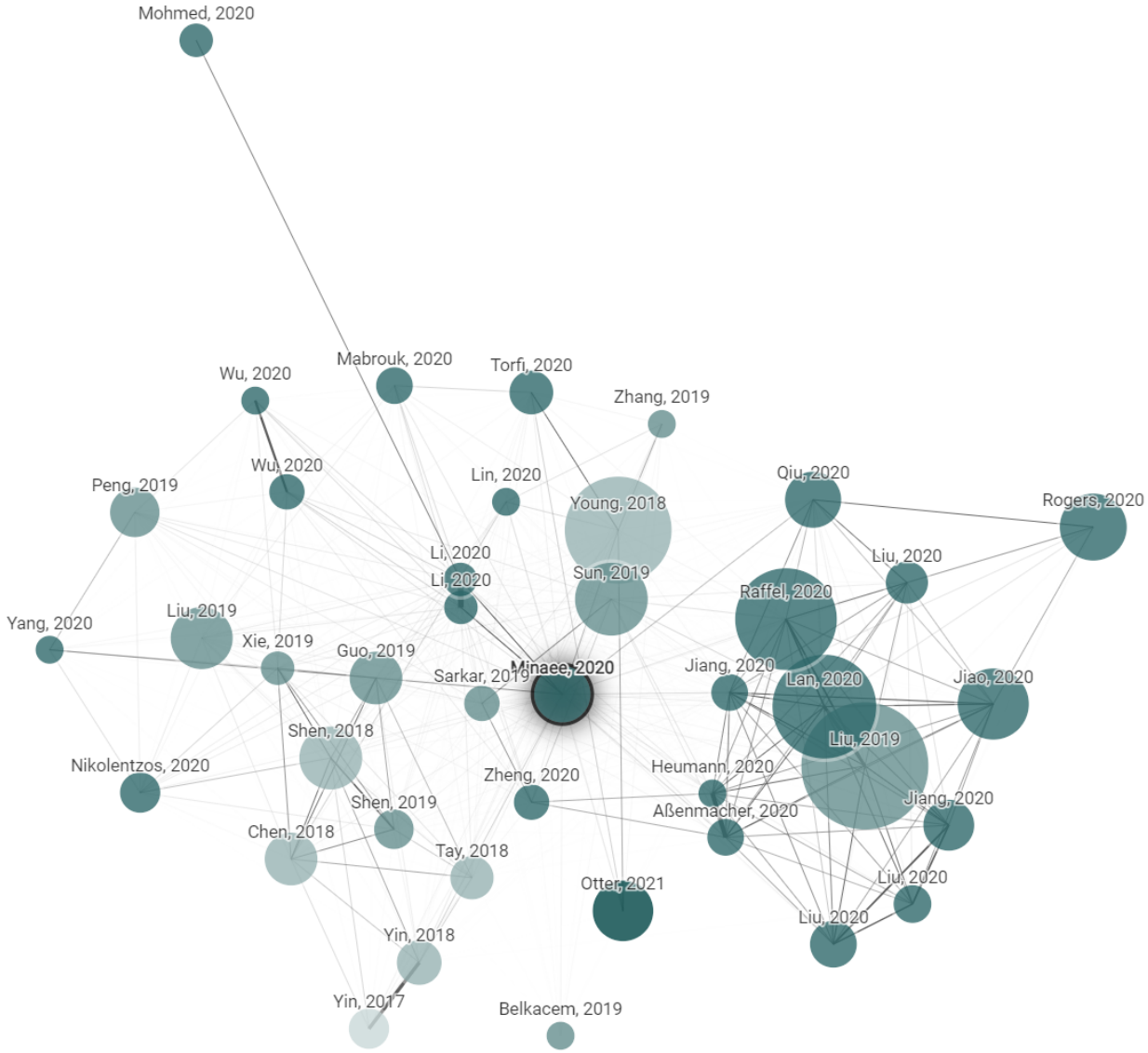


Figura 4.1: Grafo gerado pelo aplicativo *Connected Papers* relativo ao trabalho de revisão sistemática de classificação de textos realizado por Minaee et al. [95].

A rede utilizada pode ser descrita como:

$$BL(200, ?, sum) \rightarrow FL \rightarrow D(6, ?, relu)$$

Onde $BL(u, dr, co)$[3] é uma camada BiLSTM com u nós, com dr taxa de *dropout* e com a junção das camadas com a função co e FL[4] é uma camada de achatamento das entradas (*flattener*).

O trabalho alcançou um F1-*score* que variou de 73% a 93% com valor médio de 84%. Ainda em estudo prévio na mesma base [51], foi implementada rede neural convolucional (CNN) com três camadas em paralelo, mas que apresentou resultados significativamente piores.

Pesquisa desenvolvida por Shen et al. [112] propõe um modelo de sentenças (denominado Bi-BloSAN) baseado em uma rede bidirecional com *self-attention* (auto-atenção) livre de camadas recorrentes e convolucionais com o intuito de ser de execução rápida e de baixo consumo de memória. Na tarefa de classificação de sentenças atingiu resultados similares a vários classificadores de referência, sendo mais eficiente (em tempo de processamento para parâmetros idênticos) que redes recorrentes bidirecionais como Bi-LSTM e Bi-GRU.

4.2.3 Modelos de Linguagem - *Language Models*

Apesar de não ter sido abordado na Seção 4 os modelos de linguagem têm sido utilizado para inúmeras tarefas e tem se tornado maiores e mais complexos nos últimos anos[5].

Os modelos de linguagem modernos tem como grande dificuldade de implantação devido ao grande número de parâmetros (na ordem de grandeza de centenas de bilhões) [33]. Uma forma de contornar isto é usando versões "destiladas" dos modelos [109].

Bert

O modelo BERT (*Bidirectional Encoder Representations from Transformers*) foi implementado por Sun et al. [121] para classificação de texto e baseado em uma arquitetura com um *encoder* com 12 blocos de *transformer*, com 12 *heads* de auto-atenção e tamanho oculto de 768. O modelo é limitado a uma entrada de 512 *tokens* [57]. Os modelos pré-treinados do BERT tem mais de 100 milhões de parâmetros sendo o modelo "Base" com 110 milhões e o "Large" com 340 milhões [57].

[3] Não foi mostrado no artigo o valor definido de *dropout*.

[4] A camada de *flattener* não estava descrita, mas é necessária para gerar uma correta saída para a camada densa subsequente.

[5] https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, visitado em 20 de fevereiro de 2021

No estudo de Sun et al. [121] foram classificadas 8 bases conhecidas de classificação de textos em inglês e chinês). A implementação foi baseada em um ajuste fino (*fine tunning*) do modelo original de Devlin et al. [57]. Obteve acurácia acima de 95% exceto para a classificação Yelp[6] de cinco classes onde obteve acurácia pouco acima de 70%.

XLNet

A rede XLNet desenvolvida por Yang et al. [130] é baseada no BERT (particularmente o BERT-Large) com alguns ajustes na concepção de probabilidade, inclusão de ideias do modelo Transformer-XL [48] e aumentando a base de treinamento.

O modelo alcançou acurácia um pouco melhor que o BERT em todos os modelos de classificação de texto abordados por Sun et al. [121] e que a otimização do BERT publicada no mesmo ano RoBERTa desenvolvido por Liu et al. [88].

GPT-3

Este modelo de linguagem [33], cuja sigla significa *Generative Pre-trained Transformer* 3, foi desenvolvido pela OpenAI e apesar dos alardeados resultados e do tamanho massivo (até 175 bilhões de parâmetros) tem se mantido[7] atrás de uma Interface de Programação de Aplicação (API) e, desta forma, não permite que técnicas de *fine tunning* ou destilamento sejam aplicadas. Segundo o site oficial isso se dá para que sejam evitados usos maliciosos do modelo e para que ele possa ser comercializado.

ULMFit

De Araujo et al. [52] fizeram um trabalho de classificação sobre a origem (órgão público responsável) de publicações do diário oficial do Distrito Federal, desta forma tratasse de uma classificação de múltiplas classes. NComo referência de estado da arte, Araújo *et al.* utilizaram o modelo ULMFiT (*Universal language model fine-tuning*) desenvolvido por Howard and Ruder [76].

Neste estudo foi realizado confronto de SVM e UMLFiT com recursos de BOW (*Bag of Words*) e TF-IDF. Conclui que modelos SVM com média ponderada de F1-*score* de 89,17% ainda é competitivo com métodos mais modernos como o ULMFit com 89,74% no mesmo critério.

Este capítulo abordou os trabalhos atuais relacionados ao desenvolvimento deste trabalho. O capítulo seguinte apresenta a metodologia que foi seguida para se alcançar os resultados esperados.

[6] https://www.yelp.com/dataset, visitado em 20 de fevereiro de 2021
[7] https://openai.com/blog/openai-api/

5

Metodologia Proposta

"You can fight
Without ever winning
But never ever win
Without a fight ."

– Neil E. Peart / Gary L. Weinrib / Alex Zivojinovich

A Figura 5.1 é um modelo **inicial** da arquitetura proposta do sistema *Deep Vacuity* que se propõe a ser o modelo de soluções computacionais e de dados para o problema estudado neste trabalho. Esta estrutura apresentada é a inspiração da metodologia proposta aqui apresentada a seguir.

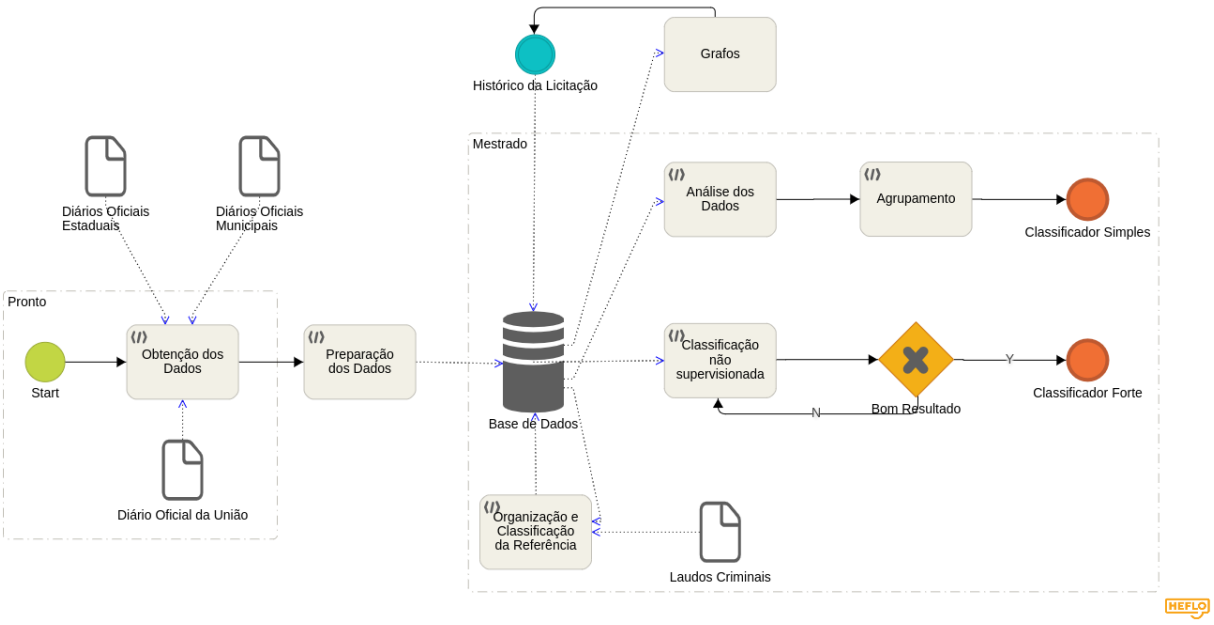


Figura 5.1: Proposta de Arquitetura do Sistema *Deep Vacuity*. [Figura do Autor]

Na Figura 5.1 é destacada uma seção marcada como "Pronto" referente a etapa do *crawler* elaborada no trabalho de [65]. No trabalho atual se obteve os recursos computacionais e de rede necessários para permitir, com pequenas adaptações no código, a obtenção e processamento dos dados públicos desde janeiro de 1998.

A extração de dados dos diários oficiais estaduais e municipais exige uma abordagem diferenciada devido a diversidade de apresentação e disponibilização dos dados e está sendo desenvolvida de forma aberta pelo projeto Querido Diário [27] e, assim, sua utilização e adaptação será deixada para um trabalho futuro.

Da mesma forma, a vinculação de todas as interligações entre as mais diversas publicações do DOU, como por exemplo a vinculação das etapas da licitação (edital, julgamento, contratos, aditivos etc.) ou a vinculação da relação entre órgãos públicos, pessoas físicas e empresas privadas, poderia adicionar valiosas informações ao classificador mas esta etapa ainda não será atacada na metodologia aqui apresentada.

As demais etapas, da arquitetura apresentada em 5.1 estão sendo desenvolvidas em outros trabalhos de pesquisa, estas realizadas por demais membros da equipe do projeto, sendo citada neste trabalho somente estritamente quando necessário. Isso foi adotado como premissa, pois alguns módulos ainda estão sob sigilo de projeto e também com intuito de manter o foco no problema de detecção para risco de conluio em publicações do DOU e adjacências.

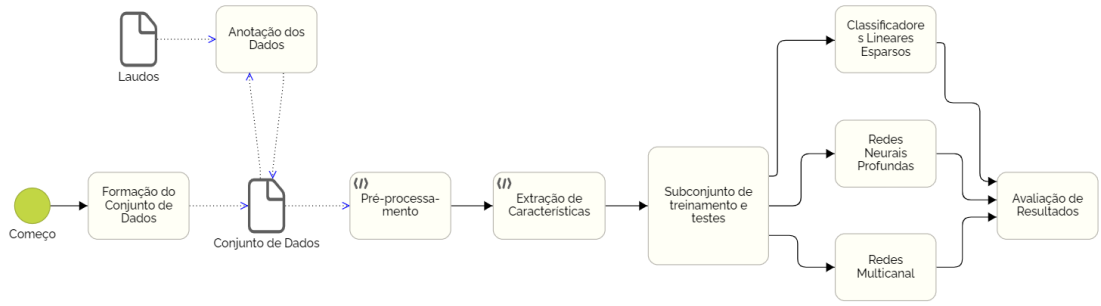


Figura 5.2: Etapas da metodologia.

Seguindo as etapas do fluxo da metodologia proposta (Figura 5.2), e tendo em mente os objetivos propostos no Capítulo 1, seguiu-se os passos de formação do conjunto de dados de publicações da Seção 3 do Diário Oficial da União, focando-se no período relativo aos dados das análises disponíveis nos Laudos da Polícia Federal, ou seja, a partir de 1998 (Vide Seção 5.1).

A partir daí, foi necessária a anotação dos dados onde foram utilizadas técnicas automatizadas e manuais para se extrair os dados das publicações e identificar aquelas vinculadas à identificação de risco de fraude dos Laudos da Polícia Federal (Seção 5.1.1).

De posse do conjunto de dados formado e anotado, procedeu-se aos passos de classificação de textos, partindo do pré-processamento, a extração de características, a separação dos conjuntos com validação cruzada e a classificação por modelos lineares esparsos e redes neurais (Seção 5.2 e seguintes).

5.1 Formação do Conjunto de Dados

O conjunto de dados proposto é uma grande coleção de fragmentos de textos extraídos do Diário Oficial da União, ou seja, de publicações todas de origem pública e aberta, disponíveis no repositório institucional da Imprensa Nacional[1].

A Lei das Licitações nº 8.666/1993 e a Lei dos Pregões nº 10.520/2002 [17, 18] obrigam todos os órgãos públicos a seguirem um número exigente de regras para qualquer contratação, sendo especificamente mais exigente no caso de obras públicas. Por força destas leis e pelo princípio constitucional da publicidade, todos os processos licitatórios realizados pela administração pública federal têm obrigação de publicar seus avisos e extratos neste jornal, consequentemente, pode-se considerar estes dados acessíveis e confiáveis para a formação do conjunto de dados.

Existem várias formas de se obter dados sobre licitações públicas como sistemas de licitação (ComprasNet, Licitacoes-e, etc.), portais de transparência, entre outros. Entretanto, no que se refere a obras públicas, as informações estão espalhadas em muitos *sites* e órgãos diferentes e disponibilizado das mais diversas maneiras e formatos, tabelas e documentos. Assim, outra característica indispensável dos dados deve ser sua completude [128] e isto só pode ser alcançado sabendo que no Brasil o *DOU* [29] é o jornal onde todos os atos que envolvam a União são obrigatoriamente publicados.

Apesar do pequeno nível de detalhes das publicações do Diário Oficial da União os dados apresentados são bastante consistentes trazendo as informações vitais acerca das licitações e contratos: valores, tipos, localização, partes envolvidas, objeto etc. As publicações listam, sem exceção, todas as obras que contam com financiamento direto da administração federal. Isso traz aos dados a característica de confiabilidade [128] para serem utilizados tanto para uma pesquisa acadêmica quanto para uma investigação criminal.

[1] Acesso ao repositório da Imprensa Nacional no endereço web: https://www.gov.br/imprensanacional/pt-br.

O conjunto de dados, como afirmado anteriormente, é formado por publicações como por exemplo cada um dos extratos mostrados nesta página do DOU na Figura 5.3.

ISSN 1677-7069

DIÁRIO OFICIAL DA UNIÃO

REPÚBLICA FEDERATIVA DO BRASIL • IMPRENSA NACIONAL

Ano CLVIII Nº 102 — Brasília - DF, sexta-feira, 29 de maio de 2020 — SEÇÃO 3

Sumário

..... Esta edição completa do DOU é composta de 211 páginas .....

Presidência da República

CASA CIVIL

INSTITUTO NACIONAL DE TECNOLOGIA DA INFORMAÇÃO

RESULTADO DE JULGAMENTO
PREGÃO Nº 2/2020

O pregoeiro oficial do Instituto Nacional de Tecnologia da Informação - ITI comunica o resultado do Pregão Eletrônico nº 02/2020 com o objeto de contratação de empresa para prestação dos serviços de recepcionistas, declarando vencedora a empresa JDR Servies Ltda., CNPJ 22.463.530/0001-09, com o valor global de R$ 106.608,96.

ORNEL COSTA DE AZEVEDO

(SIDEC - 28/05/2020) 243001-24008-2020NE800001

SECRETARIA-GERAL

SECRETARIA ESPECIAL DE ADMINISTRAÇÃO

EXTRATO DE CONTRATO Nº 28/2020 - UASG 110001

Nº Processo: 00087000198202071.
DISPENSA Nº 19/2020. Contratante: PRESIDENCIA DA REPUBLICA -.CNPJ Contratado: 02744332000106. Contratado : OPCAO COMERCIO E SERVICOS DE -SINALIZACAO LTDA. Objeto: Fornecimento sob demanda de placas comuns e de representaçao para veiculos. Fundamento Legal: LEI 8.666/93 . Vigência: 28/05/2020 a 31/12/2020. Valor Total: R$13.634,00. Fonte: 100000000 - 2020NE801055. Data de Assinatura: 28/05/2020.

(SICON - 28/05/2020) 110001-00001-2020NE110001

EXTRATO DE CONTRATO Nº 30/2020 - UASG 110001

Nº Processo: 00088000351302050.
DISPENSA Nº 28/2020. Contratante: PRESIDENCIA DA REPUBLICA -.CNPJ Contratado: 00056958000102. Contratado : IDEALLE EDITORA E PUBLICIDADE LTDA.Objeto: Prestaçao de servicos tecnicos especializados de revisao e de diagramaçao de texto sob demanda. Fundamento Legal: LEI 8.666/93 . Vigência: 27/05/2020 a 27/05/2021. Valor Total: R$14.000,00. Fonte: 100000000 - 2020NE801115. Data de Assinatura: 27/05/2020.

(SICON - 28/05/2020) 110001-00001-2020NE110001

ADVOCACIA-GERAL DA UNIÃO

SECRETARIA-GERAL DE ADMINISTRAÇÃO

SUPERINTENDÊNCIA DE ADMINISTRAÇÃO NO DISTRITO FEDERAL

EXTRATO DE TERMO ADITIVO Nº 2/2020 - UASG 110161

Número do Contrato: 23/2018.
Nº Processo: 00676000302018819.
PREGÃO SSPP Nº 9/2018. Contratante: ADVOCACIA GERAL DA UNIAO -.CNPJ Contratado: 03056608000126. Contratado : J & M COMERCIO E SERVICOS DE -TELECOMUNICACOES E INFORMA. Objeto: Prorrogar a vigência do contrato inicial por mais 12 (doze) meses. Fundamento Legal: Art. 57, inciso II da Lei 8.666/93. Vigência: 07/05/2020 a 07/05/2021. Valor Total: R$55.741,92. Fonte: 100000000 - 2020NE800044. Data de Assinatura: 06/05/2020.

(SICON - 28/05/2020)

SUPERINTENDÊNCIA DE ADMINISTRAÇÃO NO RIO DE JANEIRO

AVISO DE LICITAÇÃO
PREGÃO ELETRÔNICO Nº 5/2020 - UASG 110102

Nº Processo: 00592000342201925. Objeto: A presente contratação objetiva suprir as necessidades de aquisição de Ar condicionado do tipo Self Contained com Condensação a Água gelada, Dutado, de capacidade térmica de 10 (Dez) TR para SAD/RJ e para suas unidades circunscritas na cidade do Rio de Janeiro, pelo período de 1 (um) ano, em substituição aos aparelhos cujas vidas úteis já estejam comprometidas, conforme levantamento realizado, durante a vigência do Registro de Preços.. Total de Itens Licitados: 1. Edital: 29/05/2020 das 09h00 às 12h00 e das 13h00 às 17h00. Endereço: Rua Rodrigo Silva Nº26/18º Andar, Centro - Rio de Janeiro/RJ ou www.comprasgovernamentais.gov.br/edital/110102-5-00005-2020. Entrega das Propostas: a partir de 29/05/2020 às 09h00 no site www.comprasnet.gov.br. Abertura das Propostas: 10/06/2020 às 14h00 no site www.comprasnet.gov.br. Informações Gerais: O Edital deverá ser baixado junto ao sitio www.comprasgovernamentais.gov.br.

ANA CAROLINA SBARDELOTTI
Pregoeira

(SIASGnet - 27/05/2020) 110161-00001-2020NE000096

Ministério da Agricultura, Pecuária e Abastecimento

SECRETARIA EXECUTIVA

SUPERINTENDÊNCIA FEDERAL DE AGRICULTURA, PECUÁRIA E ABASTECIMENTO DE MINAS GERAIS

AVISO DE LICITAÇÃO
PREGÃO ELETRÔNICO Nº 3/2020 - UASG 130056

Nº Processo: 21028002933202055. Objeto: Contratação de serviços de vigilância patrimonial, vigia patrimonial e portaria para a sede da Superintendência Federal de Agricultura, Pecuária e Abastecimento no Estado de Minas Gerais (SFA-MG).. Total de Itens Licitados: 4. Edital: 29/05/2020 das 08h00 às 12h00 e das 13h00 às 17h00. Endereço: Avenida Raja Gabaglia, 245, Cidade Jardim Belo Horizonte/MG ou www.comprasgovernamentais.gov.br/edital/130056-5-00003-2020. Entrega das Propostas: a partir de 29/05/2020 às 08h00 no site www.comprasnet.gov.br. Abertura das Propostas: 10/06/2020 às 08h30 no site www.comprasnet.gov.br. Informações Gerais: .

MARCILIO DE SOUSA MAGALHAES
Ordenador de Despesas

(SIASGnet - 27/05/2020) 130056-00001-2020NE800098

SUPERINTENDÊNCIA FEDERAL DE AGRICULTURA, PECUÁRIA E ABASTECIMENTO DE SERGIPE

RETIFICAÇÃO

No Extrato de Termo Aditivo Nº 1/2020 publicado no D.O. de 27/05/2020 , Seção 3, Pág. 2. Onde se lê: Vigência: 23/05/2020 a 05/02/2021 Leia-se : Vigência: 23/05/2020 a 05/05/2021

(SICON - 28/05/2020)

COMISSÃO EXECUTIVA DO PLANO DA LAVOURA CACAUEIRA

EXTRATO DE TERMO ADITIVO Nº 2/2020

A Comissão Executiva do Plano da Lavoura Cacaueira - CEPLAC e a Empresa ÓTIMA ENGENHARIA EIRELI - EPP, celebram e ajustam o 6º Termo Aditivo decorrente do Termo de Contrato nº 10/2015. Objeto: As alterações das Cláusulas Segunda, Terceira e Quarta, do Contrato de Prestação de Serviços nº 10/2015, conforme Processo de Licitação SEI nº 21084.000696/2014-75 e 21084.000274/2017-42. Fundamento Legal: Lei nº 8.666 de 21 de junho de 1993 e suas alterações. Vigência: 06 (seis) meses, a partir da data de sua assinatura. Assinatura: 27/05/2020. Vigência: 27 de maio de 2020, até 27 de novembro de 2020. Valor R$290.435,76. Signatários: Sr. Waldo Luiz Coutinho Britto, como Representante da Contratante/CEPLAC e Sr. ORLEY SARAIVA MARIANO, como representante da Contratada.

EXTRATO DE TERMO ADITIVO Nº 1/2020

A Comissão Executiva do Plano da Lavoura Cacaueira - CEPLAC e a Empresa O FAZENDÃO SERVIÇOS AGRÍCOLAS LTDA, celebram e ajustam o 6º Termo Aditivo decorrente do Termo de Contrato nº 02/2015. Objeto: As alterações das Cláusulas Segunda, Terceira e Quarta, do Contrato de Prestação de Serviços nº 02/2015, conforme Processo de Licitação SEI nº 21084.000696/2014-75 e 21084.000274/2017-42. Fundamento Legal: Lei nº 8.666 de 21 de junho de 1993 e suas alterações. Vigência: 06 (seis) meses, a partir da data de sua assinatura. Assinatura: 27/05/2020. Vigência: 27 de maio de 2020, até 27 de novembro de 2020. Valor R$925.921,80. Signatários: Sr. Waldo Luiz Coutinho Britto, como Representante da Contratante/CEPLAC e Sr. Márcio da Silva Maciel Monteiro, como representante da Contratada.

SECRETARIA DE DEFESA AGROPECUÁRIA

DEPARTAMENTO DE SERVIÇOS TÉCNICOS

COORDENAÇÃO-GERAL DE APOIO LABORATORIAL

LABORATÓRIO FEDERAL DE DEFESA AGROPECUÁRIA EM MINAS GERAIS

EXTRATO DE DISPENSA DE LICITAÇÃO Nº 8/2020 - UASG 130058

Nº Processo: 21181000266202001 . Objeto: Aquisição de insumos para o método de determinação de glicósi em cerveja. Total de Itens Licitados: 00013. Fundamento Legal: Art. 24º, Inciso IV da Lei nº 8.666 de 21/06/1993.. Justificativa: Urgência de atendimento de situação que possa ocasionar prejuízo ou comprometer a segurança de pessoas. Declaração de Dispensa em 27/05/2020. RENATO LUIZ GONCALVES ARAUJO. Chefe da Divisão Administrativa. Ratificação em 28/05/2020. ANDREA MELO GARCIA DE OLIVEIRA. Coordenadora do Lfda/mg. Valor Global: R$ 38.997,28. CNPJ CONTRATADA : 01.530.501/0002-23 CARVALHAES PRODUTOS PARA LABORATORIO LTDA. Valor: R$ 13.632,78. CNPJ CONTRATADA : 06.880.842/0001-61 LAS DO BRASIL COMERCIO DE PRODUTOS ANALITICOS E LABORATORIAIS LTDA. Valor: R$ 17.512,50. CNPJ CONTRATADA : 43.887.033/0001-08 LABORGLAS IND E COM DE MATERIAIS PARA LABORATORIO LTDA. Valor: R$ 1.363,00. CNPJ CONTRATADA : 68.337.658/0001-27 SIGMA-ALDRICH BRASIL LTDA. Valor: R$ 6.489,00

(SIDEC - 28/05/2020) 130058-00001-2020NE800049

RESULTADO DE JULGAMENTO
PREGÃO ELETRÔNICO Nº 12/2020

O LFDA/MG torna público que a licitação na modalidade Pregão Eletrônico nº 12/2020, cujo objeto é a aquisição de óleo combustível para atender às necessidades do Laboratório Federal de Defesa Agropecuária - LFDA/MG, foi homologada. Após abertura,

Figura 5.3: Amostra de página atual do Diário Oficial da União.

Os Diários Oficiais da União na Seção 3 podem ser obtidos de janeiro de 1998 a janeiro de 2018 em formato PDF (*Portable Document File*) e posteriormente convertidos para formato de texto. Desde fevereiro de 2018, as publicações estão disponíveis em formato XML (*Extensible Markup Language*) e organizadas com campos para órgão público, tipo de documento etc., mas sem um campo específico para dados de compras, por exemplo tipo de licitação, prazos, valores e escopo. Uma publicação como da Figura 5.4 será salva no conjunto de dados em formato JSON (*JavaScript Object Notation*) , como apresentado na Tabela 5.1:

Tabela 5.1: Exemplo de publicação disponível em formato XML com os respectivos campos [29].

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

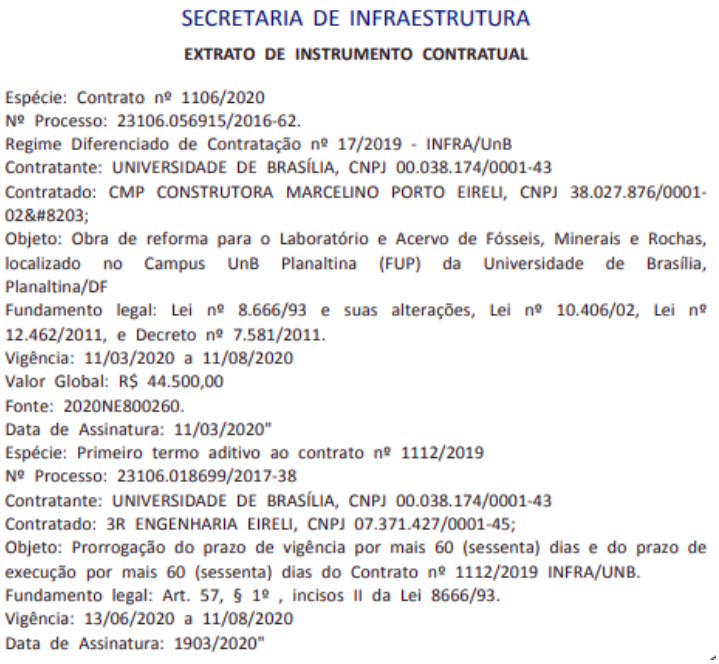
Continua na próxima página.

Tabela 5.1: continuação

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

Para publicações anteriores a fevereiro de 2018 apenas os campos "_id", "date", "indice" e "text" estão disponíveis uma vez que as demais informações são provenientes do arquivo XML. Depois de fevereiro de 2018, as tabelas são mostradas em texto com formatação HTML.

O comprimento do campo de texto varia de uma dúzia a milhares de caracteres, com títulos de subseções e nomes de assinaturas sendo os mais curtos e as editais de concursos



SECRETARIA DE INFRAESTRUTURA

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
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 11/08/2020
Valor Global: R$ 44.500,00
Fonte: 2020NE800260.
Data de Assinatura: 11/03/2020"
Espécie: Primeiro termo aditivo ao contrato nº 1112/2019
Nº Processo: 23106.018699/2017-38
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 11/08/2020
Data de Assinatura: 1903/2020"

Figura 5.4: Amostra de publicação aos Contrato nº 1106/2020 e aditivo ao contrato nº 1112/2019 da universidade de Brasília (Publicação escolhida aleatoriamente e disponível no url https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=31/03/2020&jornal=530&pagina=63.

com listas de nomes as mais longas. Embora não seja um texto estruturado, ele mantém traços individuais de ordem, pois segue tem uma maneira formal e direta de comunicação (Vide Tabela 6.1 na Seção 6.1).

5.1.1 Anotação dos Dados

A anotação dos dados públicos foi feita com a utilização de uma rede de conhecimento de especialistas forenses (Peritos Criminais da Polícia Federal) e não representam uma avaliação oficial sobre qualquer pessoa ou entidade pública e privada (vide **Aviso Legal** - pág. 5). As aquisições, contratos, trabalhos e/ou acordos foram anotados como tendo risco de fraude com base na análise de especialistas. Em sua análise são considerados múltiplos indicadores sobre data, local, tipo, peças (agência, empresas), valor, preços, execução, relação com outras publicações, e quaisquer outras informações vinculadas ao processo. Como não se levou em conta o julgamento de possíveis processos criminais ou civis não pode haver vinculação de qualquer pessoa física ou jurídica com a anotação de risco. Portanto, não são avaliados como suspeita de fraude ou não, mas como risco de fraude, e então as publicações foram marcadas como tendo $risco = 1$.

O processo de localização das publicações seguiu um processo cauteloso manual onde não seriam admitidos erros de anotação e como, nos laudos da Polícia Federal não há

menção direta ao texto das publicações do caso sob análise, o processo seguiu os seguintes passos:

1. Extração semi-automatizada com auxílio de RegEx dos dados numéricos e de identificação dos laudos:

    - Números do Convênio, Licitação, Contratos e Aditivos
    - Números de CPF e CNPJ
    - Nome dos órgãos públicos: Nomes e Siglas
    - Nome das empresas: Expressões que terminam com S.A., Ltda., EIRELI etc.
    - Datas: tanto no formato numeral quanto por extenso, transformando anos anteriores a 2000 com dois dígitos para quatro dígitos.
    - Estado da Federação: o nome do estado do "Pará" traz uma dificuldade excepcional a este tipo de processamento, pois é comum ser apresentado com erro de digitação como "Para"
    - Nome do município
    - Número de processo

2. Extração automatizada por RegEx das mesmas informações extraídas dos laudos de todas as mais de 15 milhões de publicações.

3. Vinculação das publicações aos laudos pelo número do procedimento. Neste caso os números foram reduzidos a uma forma comum, por exemplo: o número de "contrato no 574/2002" é 5742002 e o número de "Tomada de Preços Nº 472/99" é 4721999, totalizando cerca de 2 milhões de publicações (com vínculos repetidos).

4. Filtragem das publicações vinculadas por:

    - Data: a publicação deveria estar no mesmo intervalo de tempo registrado nos laudos
    - Nomes das empresas e órgão públicos: para isso foi utilizado um comparador de *strings* que considera a distância de Levenshtein para que eventuais erros de digitação ou apresentação dos nomes pudessem ser considerados. Para os órgãos públicos foi gerada lista de siglas para que fosse comparado tanto o nome quanto a sigla.
    - Em caso de a publicação ser um contrato: comparou-se o nome do município, os CNPJs, os nomes de empresas e dos órgãos públicos.
    - Em caso de convênio bastava comparar o nome do órgão conveniente, uma vez que havia coincidência dos números

- Em caso de licitação: comparou-se o nome do município e dos órgãos públicos.

Por fim, seguiu se a seleção manual final das publicações. Um caso identificado pela Polícia Federal pode gerar mais de uma publicação anotada.

Todas as outras publicações incluídas no conjunto de dados proposto foram marcadas, a princípio, como tendo *risco = 0*. Ocorre que em uma investigação criminal ou em um processo análogo de auditoria (da CGU ou do TCU) não é possível concluir que um processo licitatório é isento de fraude, ou seja, não se consegue eliminar a possibilidade de qualquer fraude de um processo, mas apenas apontar os indícios encontrados. Em outras palavras, a falta de indícios não implica em retidão do processo.

As instruções para download da base anotada "Deep-Vacuity" são encontradas na página: http://www.cic.unb.br/~fbvidal/deepvacuity/dataset/index.html.

5.2 Pré-processamento

Os passos do pré-processamento foram abordados na Seção 2.3.1 (pág. 20 e seguintes).

Uma vez que várias etapas do pré-processamento envolvem alguma perda de informação, serão realizadas testes para determinar qual a abordagem de pré-processamento do texto melhor ajuda a representar a base de dados, destacando a ocorrência das palavras.

O pré-processamento ao transformar todas as palavras em minúsculas, remover números, símbolos acentos ortográficos e *stop-words* retira informações que não afetam o processo de classificação diminuindo a dimensionalidade da vetorização e elimina muito ruído [56].

Será testada a influência da presença ou ausência de números, da realização de *stemming* e lematização e a influência de *n-gram* (com até 3 *grams*), que podem aumentar o nível de informação em uma futura extração de características ao levar para dentro da vetorização a relação entre palavras que ocorrem com frequência em conjunto [32].

Pode-se usar o texto da publicação mostrada na Tabela 5.1 para exemplificar a ação do pré-processamento:

> 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

Após o pré-processamento, o texto da publicação fica desta forma:

> extrato instrumento contratual extrato instrumento contratual especie contrato processo regime diferenciado contratacao infra unb contratante universidade brasilia cnpj contratado cmp construtora marcelino porto eireli cnpj amp objeto obra reforma para laboratorio acervo fosseis minerais rochas localizado campus unb planaltina fup universidade brasilia planaltina fundamento legal lei suas alteracoes lei lei decreto vigencia valor global fonte data assinatura especie primeiro termo aditivo contrato processo contratante universidade brasilia cnpj contratado engenharia eireli cnpj objeto prorrogacao prazo vigencia por mais sessenta dias prazo execucao por mais sessenta dias contrato infra unb fundamento legal art incisos lei vigencia data assinatura

Datas, valores financeiros, número de contrato ou de processo etc. são elementos úteis na identificação de um determinado elemento, mas não representam bem o conjunto de elementos semelhantes, uma vez que estes valores numéricos e de identificação não se repetem entre diferentes tipos de publicação, mas apenas naquelas poucas da mesma contratação. Para testar os efeitos da retirada destes elementos será modelada uma rede multicanal com dados estruturados retirados do texto (Seção 5.4.3).

5.3 Extração de Características

Como último passo antes do processo de classificação, como abordado em 2.3.2, será feita a comparação em classificadores simples da influência da utilização da vetorização utilizandose um BoW simples (*one-hot*), Word2Vec na versão Doc2Vec[2] desenvolvida por Le and Mikolov [84] e TF-IDF.

A Tabela 5.2 mostra as características dos tipos de vetorização estudadas que foram abordadas na Seção 2.3.2. Os Vocabulários serão ajustados para eliminar palavras com baixíssima frequência e se fará uma análise exploratória do melhor "valor de corte". Para o Doc2VEC, o tamanho do vetor será selecionado o maior possível de acordo com os recursos disponíveis para que haja a melhor representação possível.

Cada uma das vetorizações será gerada para um dos conjuntos de validação cruzada e feita uma primeira classificação, aqueles que tiverem melhores resultados nas classificações

[2]Os critério de treinamento do Doc2Vec, obtidos por análise exploratória, foram: tamanho do vetor $(\text{vector}_s ize) de 300, contagem mnima de palavras (min_c ount) de 2, quantidade de pocas (epochs) de 40 e usando o algurtimo de trei$ 1)

Tabela 5.2: Características dos tipos de vetorização estudadas.

Vetorização	Forma de Obtenção	Tamanho da Saída
BoW	Montagem dos vetores *One-hot*	Vocabulário (ajustado)
Word2Vec	*Embedings* (Doc2Vec)	Tamanho do vetor (300)
TF-IDF	Cálculo de Frequência	Vocabulário (ajustado)

com modelos lineares esparsos (Seção 5.4.2) serão testados em todos os conjuntos de validação cruzada e o melhor será levado para o desenvolvimento dos outros modelos.

5.4 Etapa de Classificação

A etapa de classificação desse conjunto de dados tenta emular o comportamento esperado na avaliação de um Perito Criminal sobre a possibilidade de fraude em uma determinada contratação. Especialistas entrevistados pelos autores afirmaram que valor, órgão público, tipo de empreendimento, localização, data, tipo de construção e a correlação entre essas informações costumam levar a um bom palpite sobre o risco de fraude nas licitações. Essas variáveis reunidas são os modelos de dados estruturados descritos na Seção 4.1. Os modelos que serão desenvolvidos visam emular esta capacidade usando apenas ferramentas de processamento de linguagem natural[3].

5.4.1 Subconjuntos de Treinamento e de Testes

Para lidar com o conjunto de dados desequilibrado e generalizar os modelos com sucesso, criam-se dez subconjuntos com 1.907 entradas escolhidas aleatoriamente das publicações não anotadas para equilibrar com os dados anotados. Uma maneira de classificar os dados é considerar todas as publicações escolhidas aleatoriamente como tendo *risco = 0*. Embora haja um erro nessa suposição, ela pode ser considerada tão baixa quanto a raridade[4] da classe *risco 1*. Além disso, os dez subconjuntos criados foram divididos em um subconjunto de treinamento (com validação) e um subconjunto de teste e aplicada validação cruzada de dez vezes gerando um total de uma centena de conjuntos de treinamento. A Figura 5.5 ilustra como isso foi cumprido.

[3]A configuração da máquina onde foram treinados e avaliados os modelos é:

- Processador: AMD Ryzen Threadripper 3970X 32 Core e 64 *threads*
- Memória RAM: 64 Gb
- Vídeo: Duas placas NVIDIA 2080 RTX 8 Gb

[4]Como mostrado nos trabalhos referentes a detecção de fraudes [69, 102, 111, 126], a grande maioria das transações são conduzidas sem risco, mas dada a natureza dos processos é impossível determinar a exata taxa de risco.

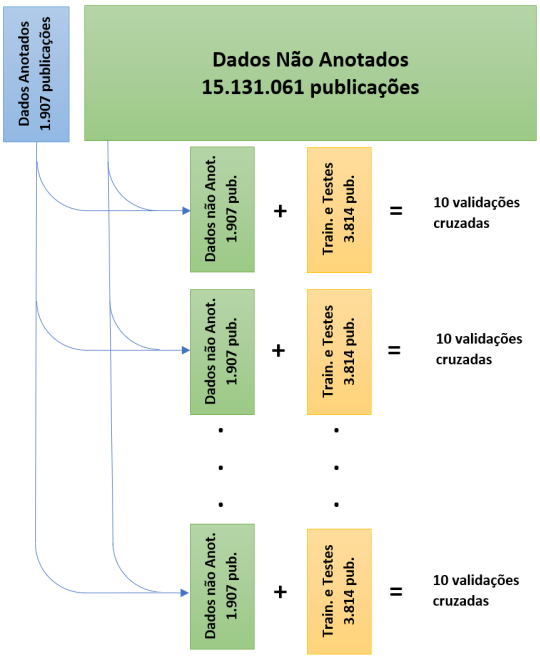


Figura 5.5: Conjuntos de treinamento e testes.

5.4.2 Comparação com Classificadores Lineares Esparsos

Para criar um referencial (*baseline*) de comparação, foi realizada classificação do conjunto de dados com uma gama de classificadores supervisionados lineares esparsos clássicos, usando modelo de extração de características TF-IDF, baseado no código de Prettenhofer et al. [103], que inclui os métodos listados da Tabela **??**:

A implementação do método classificador *Passive-Aggressive* [46] é descrito como um algoritmo online significando que, para cada resultado de predição para uma instância de uma observação sequencial, o mecanismo de predição é ajustado com base em sua correção. Um parâmetro de regularização controla esse ajuste para o método *Passive-Agressive*. Da mesma forma, a *Elastic Net* é um dos parâmetros de regularização para o método SGD [14]. [11] recomenda uma estratégia de aprendizagem online para cenários com grandes fluxos de dados, pois esse tipo de aprendizagem é baseado em cada novo evento e seus padrões.

5.4.3 Modelos de Redes Neurais Profundas

Um obstáculo atual na ciência é alcançar bons resultados usando redes neurais profundas complexas com poucos dados anotados [63], mas já é um grande avanço na investigação de fraude em licitações, uma vez que as redes neurais profundas usam um método de aprendizado supervisionado.

Tabela 5.3: Classificadores lineares esparsos baseados no código de Prettenhofer et al. [103] utilizados como base de comparação.

Método	**Observações**
Stochastic Gradient Descent (SGD)[131, 132]	Elastic Net com penalidade L1
Stochastic Gradient Descent (SGD)[131, 132]	Elastic Net com penalidade L2
SVC linear[60]	Seleção de características baseada em L1[a]
Naïve Bayes[93, 107]	Bernoulli, Complementar e Multinomial
Nearest Centroid [123]	
Passive-Aggressive [46]	
Perceptron [68]	
Random Forest [31]	
Ridge Classifier [108]	
kNN[4]	10 vizinhos

[a]A documentação do Scikit-Learn[103] expõe que: "O uso da penalização L1 conforme fornecido pelo LinearSVC produz uma solução esparsa, ou seja, apenas um subconjunto de pesos de recursos é diferente de zero e contribui para a função de decisão." [em tradução livre]

Inicialmente, para a extração de características foi utilizado o DOC2VEC [84] e, apesar de não ter trazido bons resultados, foi útil para agilizar na seleção dos modelos que seriam testados, já que na época os recursos computacionais eram mais limitados. O modelo GLOVE em português [74] também foi testado, mas não foram melhores para os resultados iniciais.

Para a extração de características do texto, a abordagem TF-IDF, já utilizada com classificadores lineares, apresentou resultados iniciais promissores. Ele foi testado e implementado com redes neurais profundas e os testes iniciais mostraram melhores resultados. O tamanho do vocabulário girava em torno de 32 mil palavras. Os modelos de rede neural profunda foram construídos usando o Tensorflow [2]. Para avaliar o conjunto de dados proposto foram escolhidos os modelos de Redes Neurais Profundas (DNN) e Redes Neurais Bidirecionais, todos apresentados no Capítulo 4 anteriormente. Dois modelos DNN foram desenvolvidos por análise exploratória, o primeiro modelo é definido como:

$$D(512, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow$$
$$D(8192, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow$$
$$D(512, 0.4, relu) \rightarrow D(256, 0.4, relu) \rightarrow D(1, 0, sigmoid)$$

Representado pela Figura 5.6, desenvolvida com a ferramenta *online* Net2Vis[5].

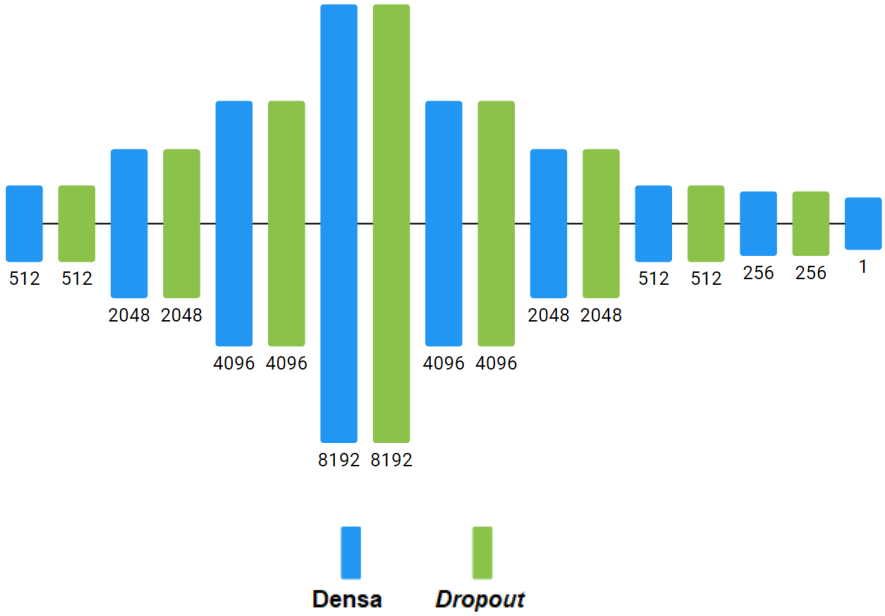


Figura 5.6: Diagrama da rede neural profunda [Figura do Autor].

Onde $D(u, dr, a)$ representa uma camada densa com u nós, com dr taxa de *dropout* e a sendo a função de ativação. A segunda rede neural profunda foi elabora da em formato de *bottleneck* (*Autoencoder*):

$$D(8192, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(256, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow$$
$$D(8192, 0.4, relu) \rightarrow D(256, 0.4, relu) \rightarrow D(1, 0, sigmoid)$$

Representado pela Figura 5.7, desenvolvida com a ferramenta *online* Net2Vis.

Muitos formatos e parâmetros do modelo de redes BiLSTM [73, 110] foram testados em conjunto de treinamento e validação, mas os melhores resultados dos conjuntos de testes foram obtidos em um primeiro momento com o seguinte:

$$BL(192, 0.25) \rightarrow BL(128, 0.25) \rightarrow BL(96, 0.25) \rightarrow TD(128, 0, relu) \rightarrow$$
$$FL \rightarrow D(1024, 0, relu) \rightarrow D(2, 0, relu)$$

[5]Disponível em https://viscom.net2vis.uni-ulm.de/

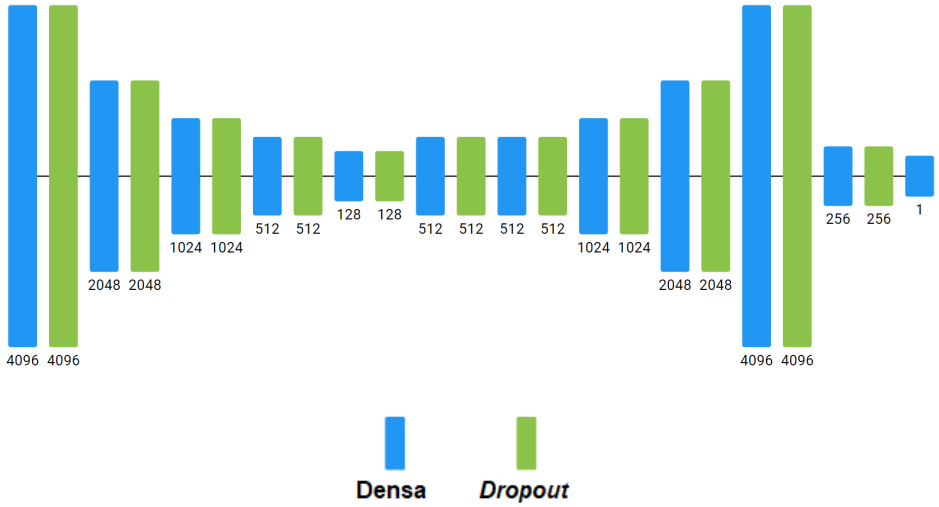


Figura 5.7: Diagrama da rede neural profunda *bottleneck*[Figura do Autor].

Representado pela Figura 5.8, desenvolvida com a ferramenta *online* Net2Vis.

Onde $BL(u, dr)$ é uma camada BiLSTM u e dr são definidos como acima, TD é uma camada densa distribuída no tempo com mesmos termos de D e FL é uma camada de achatamento das entradas (*flattener*). As saídas das camadas LSTM para frente e para trás foram combinadas por concatenação.

5.5 Redes Multicanais

Para verificar se haveria melhoria nos resultados com a conjunção de técnicas utilizadas até aqui, foi lançado mão da técnica de redes neurais multicanais já abordado na Seção 2.3.6.

Assim foi ensaiado a utilização em paralelo de duas formas de extração de características: TF-IDF e Doc2Vec tanto em redes *Autoencoders* quanto em redes profundas já utilizadas.

Da mesma forma, foram estudadas o paralelismo de redes neurais profundas e *Autoencoders*. E todos estes testes foram repetidos para as formas de fusão já citadas na Seção 2.3.6.

Uma outra abordagem foi extrair dados estruturados do texto do diário oficial de forma a gerar um conjunto mínimo de dados estruturados, criando valores escalares e binários e, assim, incluindo no modelo informações que poderiam ter sido perdidas na etapa de pré-processamento (Seção 5.2).

Estes dados extraídos são mostrados na Tabela 5.4:

Após ajustes a rede multicanal que trabalha com dois formatos de redes neurais foi testada em todo o conjunto de validação cruzada com a seguinte arquitetura:

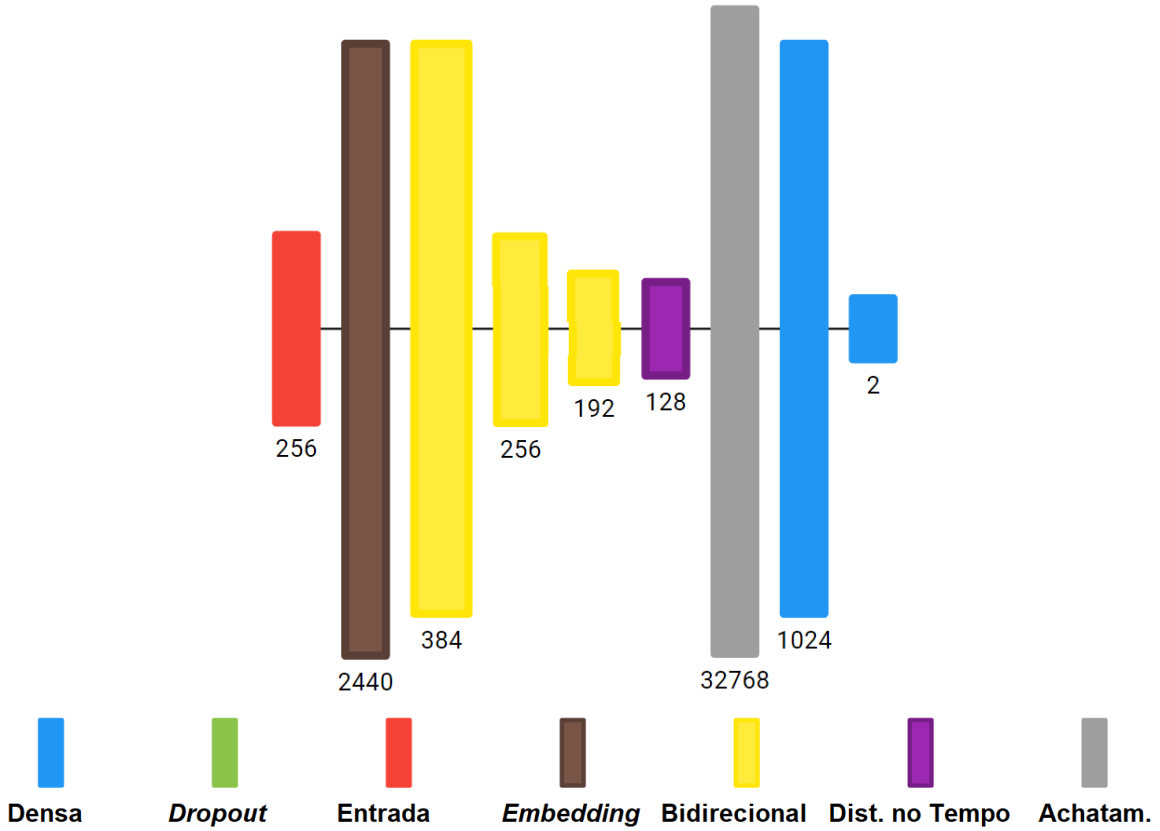


Figura 5.8: Diagrama da rede neural BiLSTM [Figura do Autor].

Tabela 5.4: Dados estruturados extraídos do texto e forma de extração

Dado	**Forma de Cálculo**
Comprimento da publicação com n caracteres	$\max\left[\frac{\log n}{10}; 1\right]$
Valor máximo presente no texto em X reais	$\max\left[\frac{\log X}{16}; 1\right]$
Número de dias corridos	desde 01/01/1998 dividido por 14.000
Licitação	0 se não e 1 se sim
Contrato	0 se não e 1 se sim
Pregão	0 se não e 1 se sim
Convite	0 se não e 1 se sim
Concorrência	0 se não e 1 se sim
Tomada de preços	0 se não e 1 se sim
Convênio	0 se não e 1 se sim
Contrato Emergencial	0 se não e 1 se sim
Dispensa	0 se não e 1 se sim
Concurso público	0 se não e 1 se sim
Leilão	0 se não e 1 se sim
CPF	0 para ausente e 1 presente
CNPJ	0 para ausente e 1 presente

$$\textbf{TF-IDF} \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(512, 0.4, relu) \rightarrow D(128, 0.4, relu) \rightarrow D(512, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(256, 0.4, relu)$$
$$\rightarrow \textbf{F(op) \& D(1, 0, sigmoid)} \leftarrow$$
$$D(256, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow D(1024, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow$$
$$D(4096, 0.4, relu) \leftarrow D(8192, 0.4, relu) \leftarrow D(4096, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow$$
$$D(1024, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow \textbf{TF-IDF}$$

Representado pela Figura 5.9, desenvolvida com a ferramenta *online* Net2Vis.

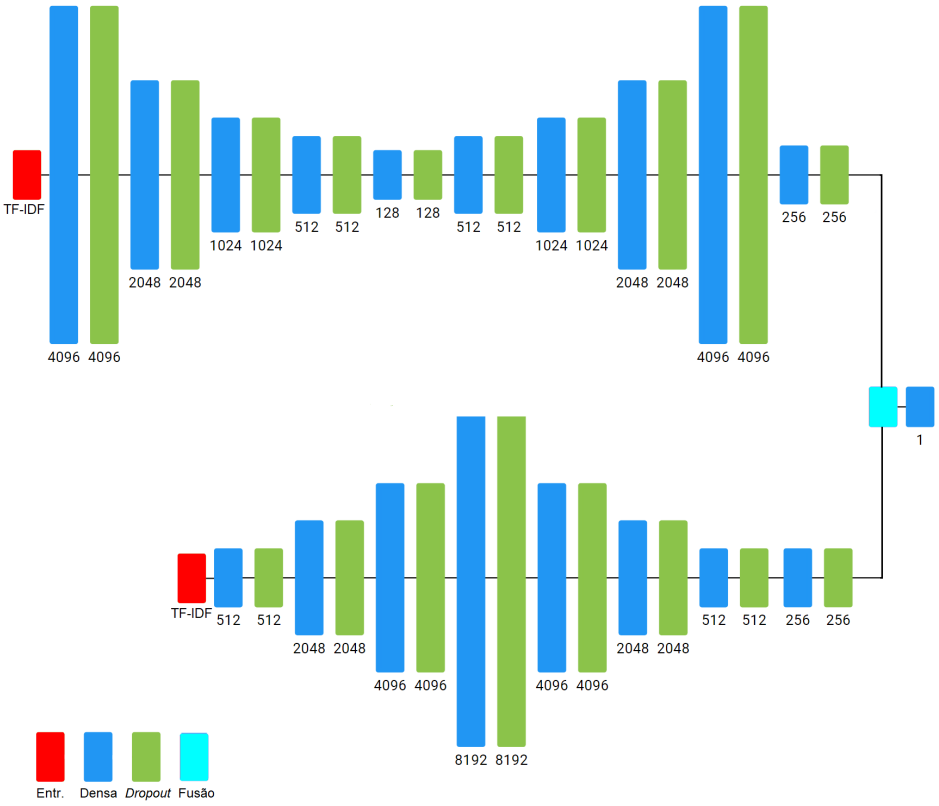


Figura 5.9: Diagrama da rede neural Multicanal com dois formatos de redes neurais e entradas idênticas TF-IDF [Figura do Autor].

Nota-se que nesta apresentação as entradas TF-IDF são duplicadas e são apresentadas no início e fim desta representação. No centro é apresentado a fusão **F(op)** das duas redes com um operador **op** seguido da camada de saída $D(1, 0, sigmoid)$.

De forma semelhante a rede *bottleneck* com a aplicação de duas formas de extração de características, assim foi implementada:

$$\textbf{TF-IDF} \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(512, 0.4, relu) \rightarrow D(128, 0.4, relu) \rightarrow D(512, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(256, 0.4, relu)$$
$$\rightarrow \textbf{F(op) \& D(1, 0, sigmoid)} \leftarrow$$
$$D(256, 0.4, relu) \leftarrow D(4096, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow$$
$$D(1024, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow D(128, 0.4, relu) \leftarrow$$
$$D(512, 0.4, relu) \leftarrow D(1024, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow$$
$$D(4096, 0.4, relu) \leftarrow \textbf{DOC2VEC}$$

Representado pela Figura 5.10, desenvolvida com a ferramenta *online* Net2Vis.

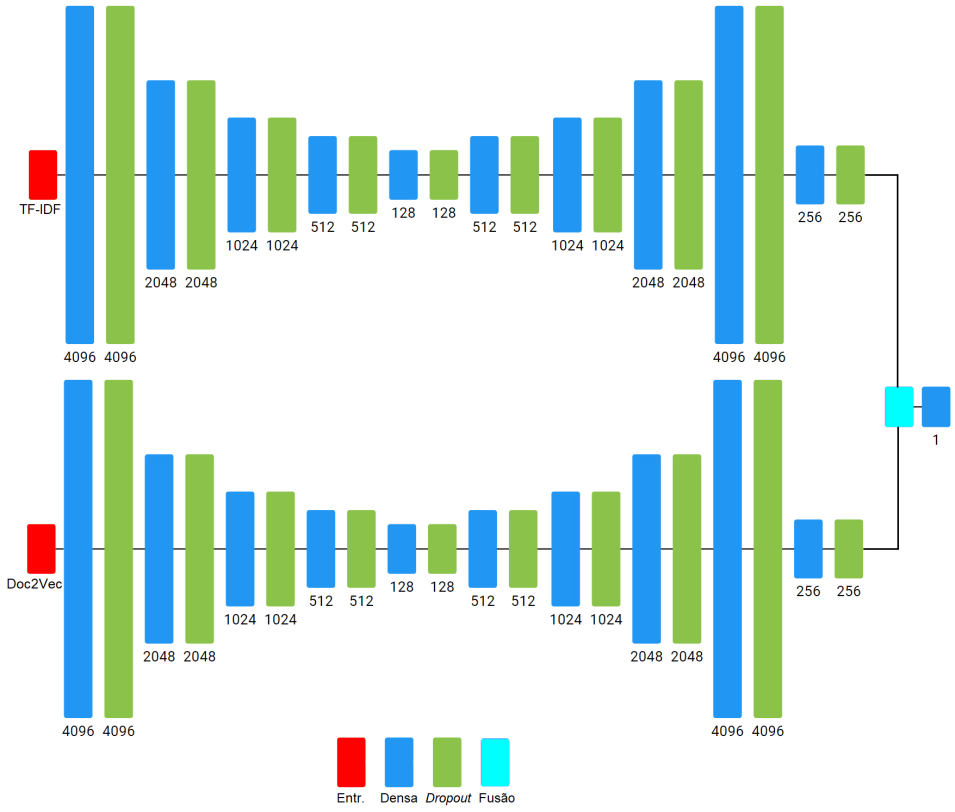


Figura 5.10: Diagrama da rede neural Multicanal com duas redes tipo *Bottleneck* e entradas TF-IDF e Doc2Vec [Figura do Autor].

E na rede profunda com estas duas extrações de características:

$$\textbf{TF-IDF} \rightarrow D(512, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow$$
$$D(4096, 0.4, relu) \rightarrow D(8132, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow$$
$$D(1024, 0.4, relu) \rightarrow D(512, 0.4, relu) \rightarrow D(256, 0.4, relu)$$
$$\rightarrow \textbf{F(op) \& D(1, 0, sigmoid)} \leftarrow$$
$$D(256, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow D(1024, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow$$
$$D(4096, 0.4, relu) \leftarrow D(8192, 0.4, relu) \leftarrow D(4096, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow$$
$$D(1024, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow \textbf{DOC2VEC}$$

Representado pela Figura 5.11, desenvolvida com a ferramenta *online* Net2Vis.

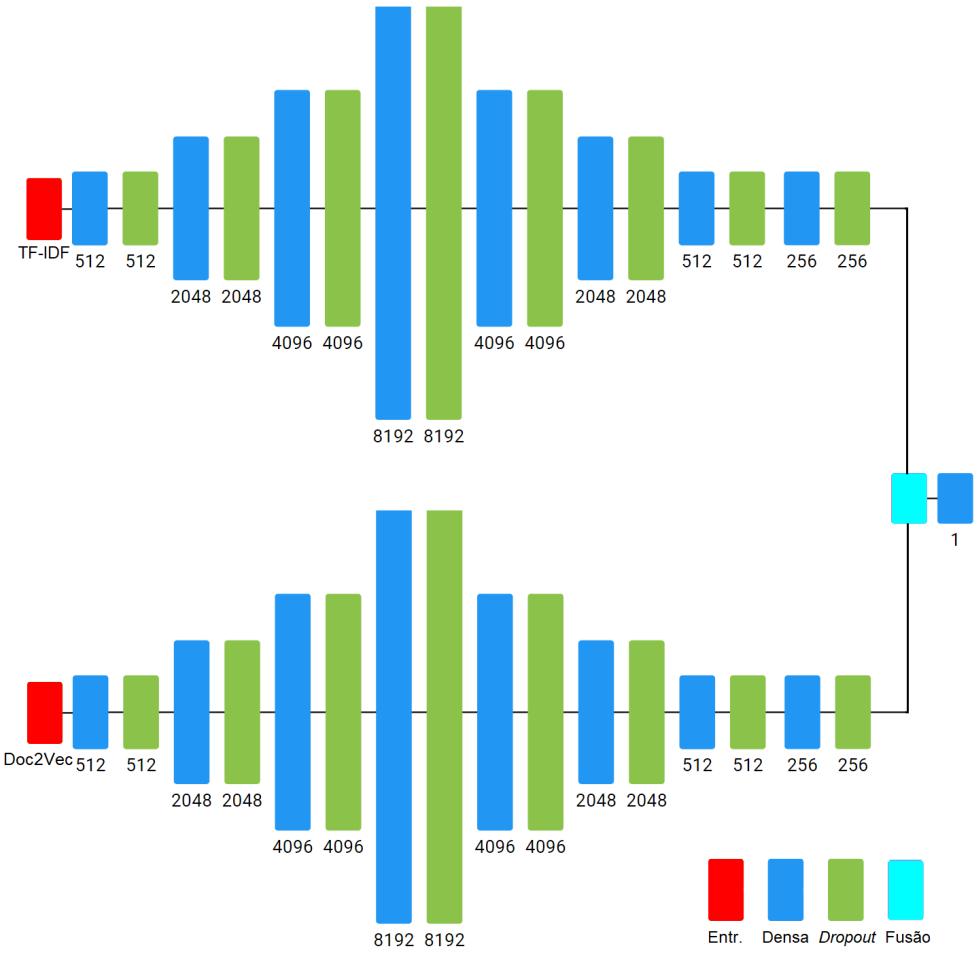


Figura 5.11: Diagrama da rede neural Multicanal com duas redes profundas e entradas TF-IDF e Doc2Vec [Figura do Autor].

Para os estudos que incluíam em paralelo os dados estruturados recém-extraídos, primeiramente se procedeu a testes de classificadores lineares, já utilizados, para verificar se existiam informações capazes de gerar classificação do conjunto de dados acerca do risco das publicações.

Após estes testes, foi desenvolvida a seguinte rede neural multicanal para classificação:

$$\textbf{TF-IDF} \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(512, 0.4, relu) \rightarrow D(128, 0.4, relu) \rightarrow D(512, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow$$
$$D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(256, 0.4, relu)$$
$$\rightarrow \textbf{F(op) \& D(1, 0, sigmoid)} \leftarrow$$
$$D(256, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow$$
$$D(64, 0.4, relu) \leftarrow D(512, 0.4, relu) \leftarrow D(2048, 0.4, relu) \leftarrow \textbf{Estruturados}$$

Representado pela Figura 5.12, desenvolvida com a ferramenta *online* Net2Vis.

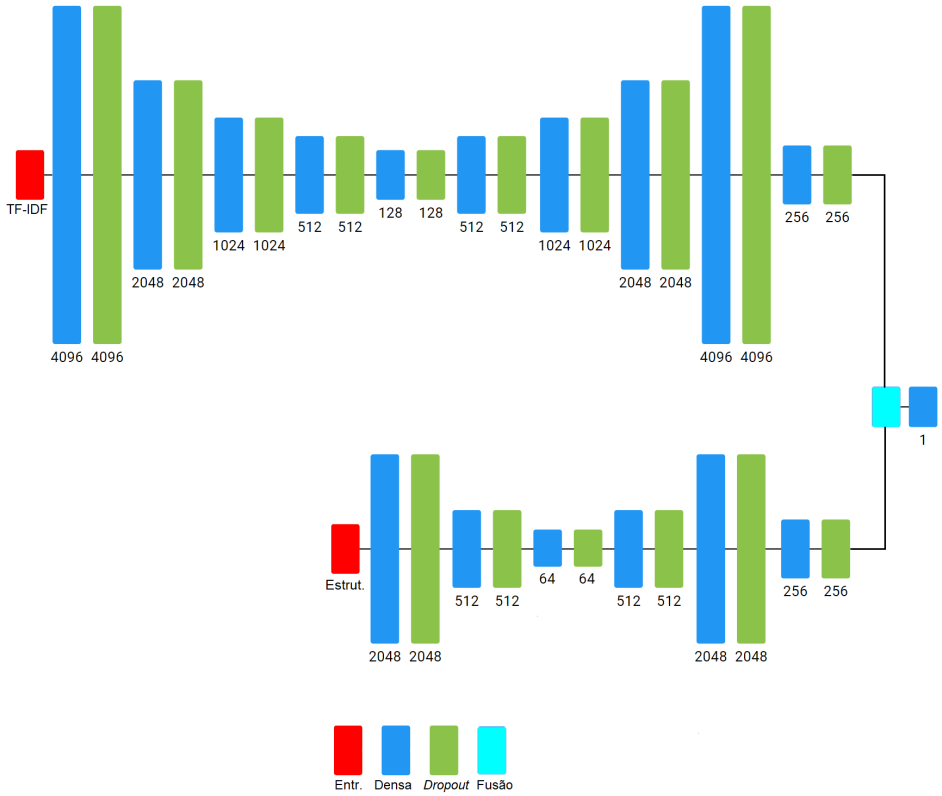


Figura 5.12: Diagrama da rede neural Multicanal com duas redes tipo *Bottleneck* e entradas TF-IDF e dados estruturados [Figura do Autor].

5.6 Avaliação de Resultados

Como abordado na Seção 2.4, se utilizou da matriz de confusão para com avaliação das taxas de Verdadeiro Negativo (TN), Verdadeiro Positivo (TP), Falso Positivo (FP) e Falso Negativo (FN) para avaliar a classificação binária.

As métricas utilizadas foram a precisão (*precision*), a sensibilidade (*recall*) e F1-*score*. Esses valores foram plotados em gráficos para facilitar a comparação entre os diferentes modelos testados.

Buscou-se modelos com um equilíbrio entre a precisão e a sensibilidade (medido pelo F1-*score*), mas dando-se preferência a modelos com melhor precisão, pois a redução da taxa de falsos positivos ajuda na redução do escopo de trabalho da investigação que pode ser gerada após a determinação do risco de fraude.

Por outro lado, não se pode selecionar um modelo com alta taxa de falsos negativos (alta sensibilidade), pois poder-se-ia não investigar casos relevantes.

Por fim, o melhor modelo foi avaliado pela métrica de AUC e, seguindo os passos da Seção 2.5, foi feito o ajuste de decisão do risco pela análise do *Equal Error Rate* de forma a permitir definir o limiar que melhor ajustaria a saída de um classificador binário.

Este capítulo abordou a metodologia que foi seguida para se alcançar os resultados esperados. O capítulo seguinte apresenta os resultados alcançados por esta metodologia.

6

Resultados

"Grandes verdades podem, somente, ser esquecidas, mas nunca podem ser falsificadas."

– G. K. Chesterton

Este capítulo aborda os resultados alcançados. Serão mostrados a formação do conjunto anotado de dados na Seção 6.1, os dados da entrada de treinamento na Seção 6.2, os resultados da classificação na Seção 6.3 e das redes multicanais na Seção 6.4 e o limiar (*Equal Error Rate*) na Seção 6.8.

6.1 Formação do Conjunto de Dados

Com a adaptação do *crawler* desenvolvido em Ferreira [65] foi possível a obtenção em formato de texto de todas as publicações da Seção 3 do Diário Oficial da União. O processo envolveu a utilização da rede privada da Polícia Federal, pois a utilização de outras redes implicava no bloqueio do IP por excesso de acessos. Obteve-se, principalmente os textos das publicações do período entre janeiro de 1998 e janeiro de 2018, que só é disponibilizado em PDF.

Como os textos a partir de fevereiro de 2018 são obtidos por meio de API e em formato XML, a complementação do banco de dados não apresentou nenhuma dificuldade, tendo publicações adicionadas até a data presente.

A Tabela 6.1 mostra as principais estatísticas do conjunto de dados. O conjunto de dados recebeu informações até o último dia útil de fevereiro de 2020 em um total de 15.132.968 entradas. Após esta data, o banco de dados é alimentado automaticamente pelo sistema *Deep Vacuity*[1] (em desenvolvimento).

[1]Sistema acessível pelo sítio web: https://deepvacuity.cic.unb.br/

Tabela 6.1: Estatísticas dos Dados baixados até fevereiro de 2020.

	Dados
Entradas	15.132.968
Comprimento Máximo	1.040.513 caracteres ou 156.703 palavras
Comprimento médio	761,0 caracteres ou 110,2 palavras

O levantamento da numeração de convênios, licitações e contratos nos documentos da Polícia Federal foi feito e conferindo manualmente cada uma das marcações. Após as filtragens semi-automatizadas descritas na Seção 5.1.1 foi feita a seleção manual final entre cerca de cinco mil publicações, restando no final 1.907 publicações anotadas como de *risco = 1*.

Assim, este total de 1.907 publicações foram anotadas o que representa $0,012\%$ do conjunto de dados. Os dados anotados abrangem publicações relacionadas a contratos de construção que variam de milhares a centenas de milhões de reais em todos os estados brasileiros.

Uma base de dados de fraude tende a ser composta por uma ampla maioria de operações ou registros (neste caso licitações) legais e uma pequena porção de fraude. E ainda, apenas uma parte desta última é de fato investigada e comprovada. Soma-se a isso o fato de o tempo exigido de um especialista para analisar completamente um processo de contratação e execução de obra pública ser elevado (variando de 3 semanas - casos simples - a 3 meses - casos complexos). Assim, o conjunto de dados é extremamente desequilibrado, resultando em uma razão de 7.935 publicações não anotadas para cada uma marcada como *risco 1*.

6.2 Dados da Etapa de Treinamento

Nesta seção serão apresentadas as configurações dos modelos de classificação que atingiram os melhores resultados na busca exploratória. Para atingir estes resultados, o os critérios de configuração foram testados em um fixo conjunto de validação cruzada, para só depois realizar os testes com os demais 99 conjuntos.

6.2.1 Dados de Treinamento dos Modelos Lineares Esparsos

Para os modelos lineares esparsos, os critérios de configuração são apresentados na Tabela 6.2.

Tabela 6.2: Configuração dos modelos lineares esparsos.

Ridge Classifier	Tolerância: 10^{-2} *solver*:SAG[a]
Perceptron	iterações: 50
Passive Aggressive	iterações 50
KNN	vizinhos: 10
Random Forest	-
SGDClassifier ElasticNet	alpha: 10^{-4}, iterações: 50
Nearest Centroid	-
MultinomialNB	alpha: 10^{-2}
BernoulliNB	alpha: 10^{-2}
ComplementNB	alpha: 10^{-1}

[a]SAG: *Stochastic Average Gradient*

6.2.2 Dados de Treinamento dos Modelos de Redes Neurais

A Tabela 6.3 mostra os dados de treinamento utilizados nas redes neurais, com o número de épocas, tamanho do *batch* e configuração da taxa de aprendizado.

Foi ensaiado o uso de *One Cycle Policy* mostrada por Smith [119], entretanto a abordagem com o método "triangular2" [118] (Vide Figura 6.1) apresentou melhores resultados.

Tabela 6.3: Critérios de configuração das redes neurais que alcançaram os melhores resultados.

Modelo	**Épocas**	***Batch Size***	***Learning Rate***
Bottleneck	100	512	10^{-5} a 10^{-3} passo de 50 épocas
Profunda	100	512	10^{-5} a 10^{-3} passo de 50 épocas
BiLSTM	200	256	10^{-5} a 10^{-3} passo de 100 épocas Canais unidos por concatenação
Multicamadas	200	256	10^{-6} a 10^{-3} passo de 100 épocas

O treinamento e ajustes dos parâmetros das redes neurais foi feito pelo acompanhamento da taxa de aprendizado (Exemplo na Figura 6.2) e pelos resultados nos dados de testes de um conjunto de validação cruzada.

6.3 Resultado da Classificação

Os dados aqui apresentados foram avaliados de acordo com as técnicas expostas na Seção 2.4 e seguindo as metodologias apresentadas na Seção 5.6 de forma a que se pudesse julgar os melhores métodos e modelos.

Para cada um dos modelos descritos nas Seções 5.4.2 e 5.4.3, foram executados cem testes de acordo com a validação cruzada (Seção 5.4.1) e foram calculados a precisão, a

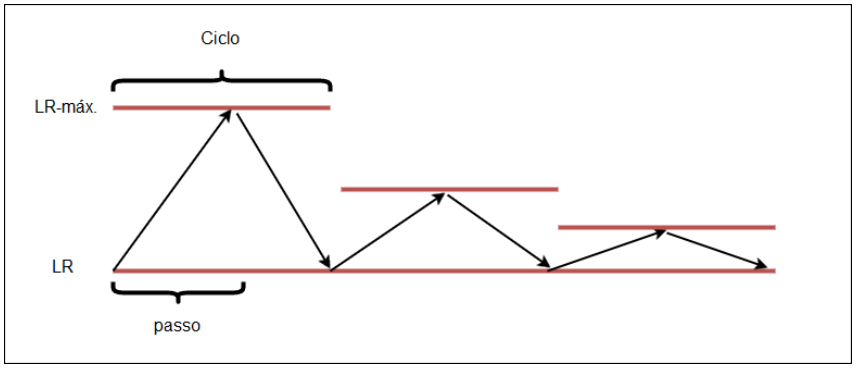


Figura 6.1: Funcionamento do método de taxa de aprendizagem cíclica tipo "triangula2" [Figura de Smith [118] disponível em https://github.com/bckenstler/CLR].

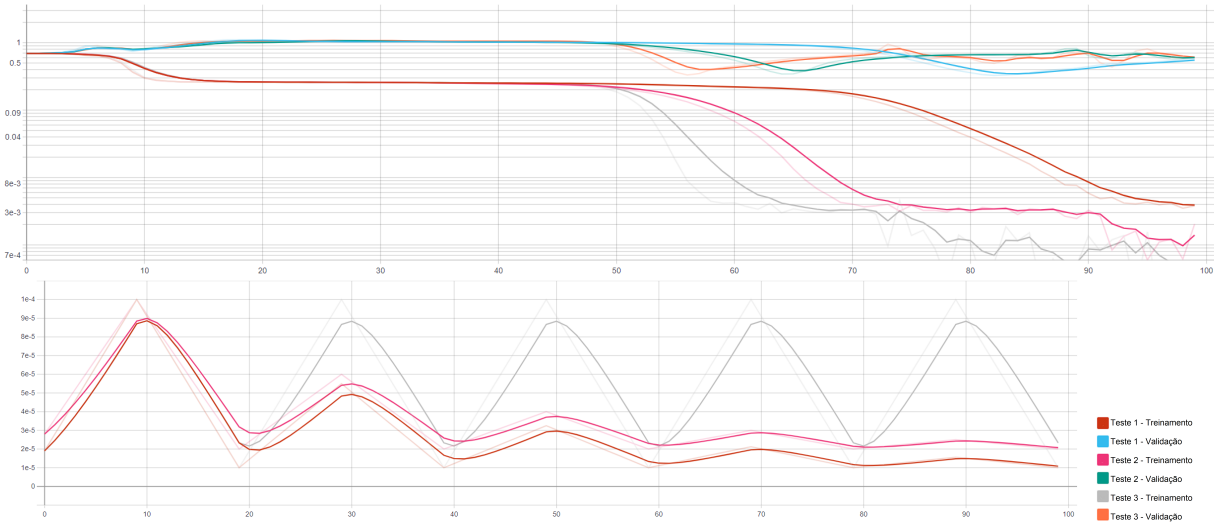


Figura 6.2: Gráfico de perda (*loss*) e de taxa de aprendizado (*learn rate*) para três treinamentos de rede neural *Botleneck* com os conjuntos de treinamento e validação [Figura do Autor - TensorBoard].

sensibilidade e a F1-*score* com o respectivo desvio padrão. Esses resultados, mostrados na Tabela 6.4, foram plotados na Figura 6.3. O retângulo ampliado indica a maioria dos classificadores lineares em um intervalo de F1 relativamente pequeno de $91,4\%$ a $93,4\%$. Pode-se observar que as redes neurais tenderam a produzir um desvio padrão mais significativo.

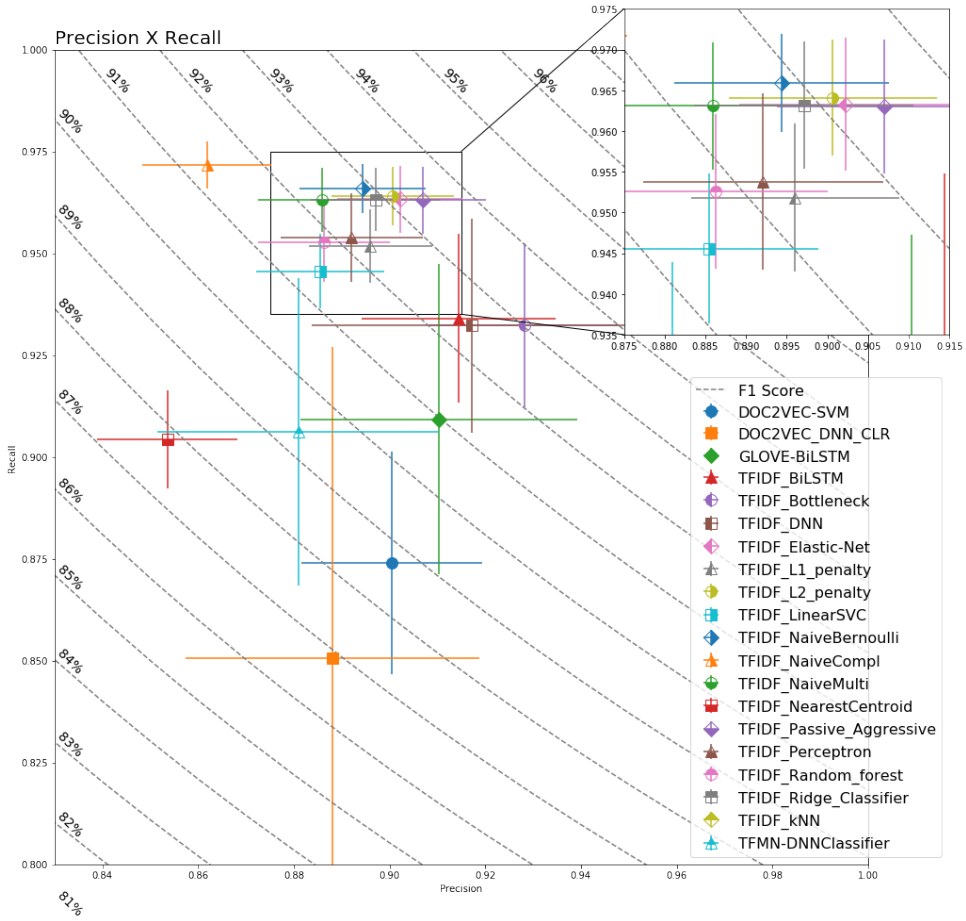


Figura 6.3: Precisão, sensibilidade e F1-*score* dos modelos.

A melhor precisão foi obtida pela rede *bottleneck* com 92,8%, a melhor Sensibilidade pela *Naïve Complement* com 97,2%, mas o melhor F1-Score foi obtido pelo *Passive-Aggressive* com 93,4%.

Embora com F1-Score menor, ambos os classificadores de redes neurais tiveram maior precisão, ou seja, classificam menos publicações como falsos positivos. E, sabendo que o modelo final pode levantar necessidade de novas investigações, e, em uma polícia de recursos finitos, essa diferença de apenas 1% equivaleria a uma média de 569 anotações de risco mensais. Desta forma, é preferível aumentar a precisão e diminuir a sensibilidade, do que o contrário.

Tabela 6.4: Resultado comparativo dos modelos treinados.

Método	**Média**		
	Precisão	**Sensibilidade**	**F1-Score**
D2V-SVM	90,1%	87,4%	88,7%
D2V-DNN	88,8%	85,1%	86,6%
GL.Bi-LSTM	91,0%	90,9%	90,9%
Bi-LSTM	91,4%	93,4%	92,4%
Bottleneck	**92,8**%	93,2%	**93,0%**
DNN	91,7%	93,2%	92,4%
Elastic-Net	90,2%	96,3%	93,2%
L1-penalty	89,6%	95,2%	92,3%
L2-penalty	90,1%	96,4%	93,1%
LinearSVC	88,6%	94,6%	91,4%
NaïveBern.	89,4%	96,6%	92,9%
NaïveCompl	86,2%	**97,2**%	91,3%
NaïveMulti	88,6%	96,3%	92,3%
NearestCen.	85,4%	90,4%	87,8%
Pass.-Agg.	90,7%	96,3%	**93,4%**
Perceptron	89,2%	95,4%	92,2%
Random-for.	88,6%	95,3%	91,8%
Ridge-Class.	89,7%	96,3%	92,9%
kNN	80,9%	95,5%	87,6%
DNNClass.	88,1%	90,6%	89,2%

6.4 Modelos Multicanais

Seguindo a metodologia da Seção 5.5, os modelos foram testados em todos os conjuntos de validação cruzada.

Foi utilizada a nomenclatura da Tabela ?? para os modelos multicanal com dois tipos de redes ou dois tipos de extração de características.

Tabela 6.5: Explicação da nomenclatura utilizada.

Sigla	Significado
BN	Rede *bottleneck*
Prof	Rede profunda
Str	Dados Estruturados
D2V	Vetorização baseada em DOC2VEC
add	Fusão dos canais por adição
avg	Fusão dos canais por média
mul	Fusão dos canais por multiplicação
sqd	Fusão dos canais pelo quadrado da diferença
abd	Diferença absoluta
con	Concatenação

6.4.1 Canais com duas vetorizações e dois modelos de redes

As redes neurais tipo *bottleneck* e profunda foram testadas em paralelo e cada uma individualmente com as vetorizações TF-IDF e Doc2Vec. Assim, os resultados são apresentados na Tabela 6.6 e na Figura 6.4.

Os modelos que incluíram a vetorização DOC2VEC (D2V) apresentaram erros nos conjuntos de testes da validação cruzada, estes modelos não convergiram para um resultado adequado, classificando todos os elementos ou como *risco 0* ou como *risco 1*, como mostrado na Tabela 6.7

A quantidade de conjuntos da validação cruzada (no conjunto de teste) que não alcançou resultado satisfatório com vetorização DOC2VEC é representativa, sendo menos relevante com a fusão de diferença quadrada (sqd) e subtração (sub). Além disso, o desvio padrão da ordem de 10% mostra que o comportamento não é constante. O melhor F1-*score* com estas configurações foi de $87,76\%$, a sensibilidade ficou alta nestes a custa de uma precisão significativamente baixa.

Com a mistura de arquitetura de redes usando apenas TF-IDF, os resultados apresentaram resultados melhores e sem erros, atingindo 91,66% de F1-*score* no caso onde a fusão das redes se deu por adição (*add*), a precisão ficou na ordem de 90% e a sensibilidade na ordem de 93%.

Tabela 6.6: Valores encontrados na validação cruzada dos modelos multicanal (excluídos os conjuntos de dados que geraram erro).

Método	**Média**		
	Precisão	**Sensibilidade**	**F1-Score**
BN-TFIDF-D2V-add	77.43%±11.80%	96.55%±3.03%	85.24%±6.39%
BN-TFIDF-D2V-avg	74.22%±12.03%	96.70%±3.41%	83.22%±6.78%
BN-TFIDF-D2V-mul	70.74%±11.92%	**98.07%±1.85%**	81.51%±7.43%
BN-TFIDF-D2V-sqd	81.19%±7.89%	96.21%±2.46%	87.76%±4.47%
BN-TFIDF-D2V-sub	76.53%±12.23%	96.42%±3.55%	84.56%±6.70%
BN-Prof-TFIDF-add	90.01%±2.35%	93.47%±2.22%	**91.66%±0.98%**
BN-Prof-TFIDF-sqd	89.42%±2.67%	93.47%±2.04%	91.35%±1.17%
BN-Prof-TFIDF-sub	**90.26%±2.92%**	92.40%±3.01%	91.23%±1.32%
Prof-TFIDF-D2V-add	74.01%±12.48%	94.77%±14.43%	81.15%±10.96%
Prof-TFIDF-D2V-avg	77.01%±11.01%	93.62%±15.28%	82.60%±13.24%
Prof-TFIDF-D2V-mul	74.87%±12.88%	93.69%±14.59%	81.04%±13.02%
Prof-TFIDF-D2V-sqd	75.31%±11.28%	94.03%±15.74%	81.48%±12.60%
Prof-TFIDF-D2V-sub	78.42%±11.83%	94.90%±9.89%	84.68%±8.28%

Tabela 6.7: Erros nos conjuntos de validação cruzada no conjunto de testes em modelos multicanais que utilizaram vetorização DOC2VEC.

Modelo	**Erros**
BN-TFIDF-D2V-add	12
BN-TFIDF-D2V-avg	10
BN-TFIDF-D2V-mul	33
BN-TFIDF-D2V-sqd	2
BN-TFIDF-D2V-sub	3
Prof-TFIDF-D2V-add	19
Prof-TFIDF-D2V-avg	30
Prof-TFIDF-D2V-mul	48
Prof-TFIDF-D2V-sqd	10
Prof-TFIDF-D2V-sub	21

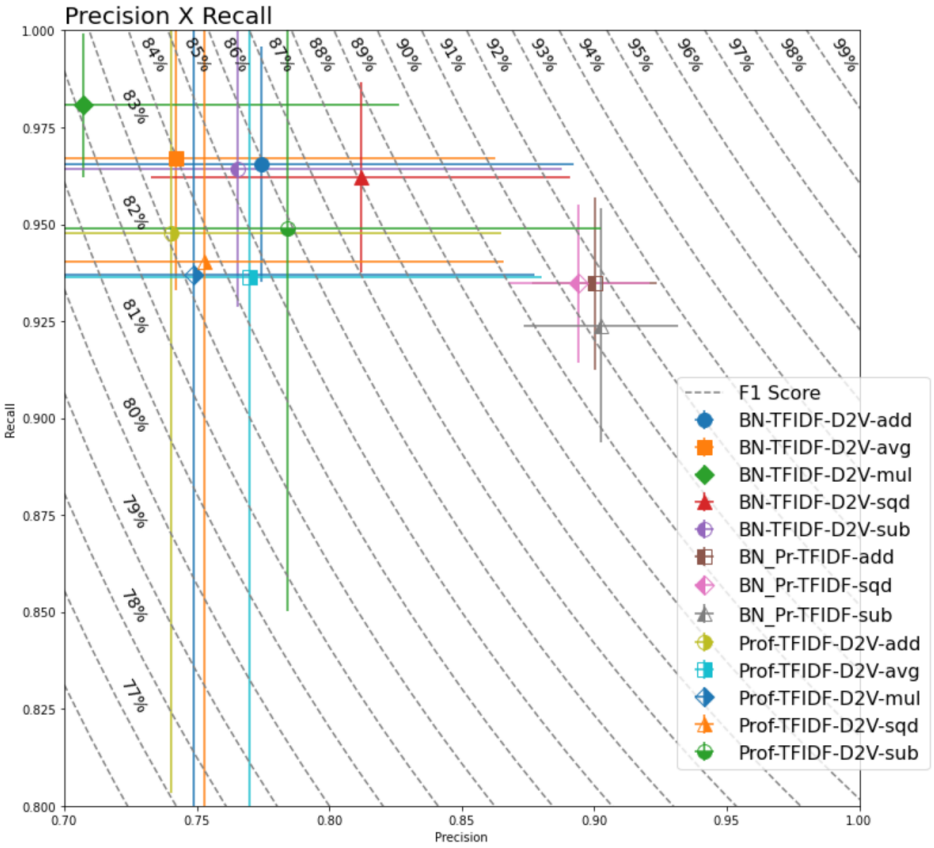


Figura 6.4: Precisão, sensibilidade e F1-*score* dos modelos multicanal.

Apesar de as redes apenas com TF-IDF apresentarem resultados melhores do que as redes com DOC2VEC, elas não superaram o obtido com as redes isoladas.

6.4.2 Redes multicanais com dados estruturados

Utilizando a mesma codificação da Tabela **??**, foram testados modelos com a extração de dados estruturados do texto.

Inicialmente, foram feitos os testes de classificação apenas com classificadores lineares para verificar se havia alguma potencial contribuição ao entendimento da base de dados, ou em outras palavras, se havia **informação** acerca do risco nestes dados.

Dos modelos testados a rede neural *SimpleANN* gerou 3 modelos (dos 100 conjuntos de validação cruzada) que não convergiram.

Os resultados destes testes são apresentados na Tabela 6.8 e na Figura 6.5.

Tabela 6.8: Valores encontrados na validação cruzada dos modelos lineares para os dados estruturados isoladamente (excluídos os conjuntos de dados que geraram erro).

Método	**Média**		
	Precisão	**Sensibilidade**	**F1-Score**
Elastic-Net	75.50%±2.49%	91.87%±6.97%	82.65%±2.36%
L1-penalty	75.38%±2.32%	92.40%±7.55%	82.75%±3.48%
L2-penalty	75.36%±2.12%	**92.47%±6.14%**	82.87%±2.24%
LinearSVC	77.58%±1.56%	86.71%±1.54%	81.88%±1.14%
NaiveBernoulli	74.97%±1.29%	89.91%±2.09%	81.75%±1.12%
NaiveCompl	73.20%±1.51%	91.16%±1.99%	81.18%±1.26%
NaiveMulti	73.19%±1.51%	91.16%±1.99%	81.18%±1.26%
NearestCentroid	74.60%±1.58%	82.91%±1.05%	78.52%±1.08%
Passive-Aggressive	76.52%±6.53%	71.92%±24.92%	70.39%±16.16%
Perceptron	75.99%±7.22%	69.30%±27.60%	67.65%±18.62%
Random-forest	**84.94%±1.32%**	90.27%±1.15%	**87.52%±0.93%**
Ridge-Classifier	77.65%±1.53%	85.85%±1.71%	81.53%±1.15%
kNN	82.22%±1.63%	88.54%±1.42%	85.24%±0.91%
simpleANN	80.74%±6.73%	63.25%±31.57%	64.32%±23.27%

Os valores apresentados com dados estruturados nos classificadores lineares mostraram, apesar dos resultados piores que os obtidos com dados vetorizados dos textos, que há possibilidade de classificação do textos usando estes dados. O melhor resultado alcançado nesta etapa foi de um f1-*score* de $87,52\%$ pelo modelo de *random-forest*.

Para estes dados foram desenvolvidas redes neurais multicanais no formato *bottleneck*, sendo que um canal ela alimentado pelos dados textuais em TF-IDF e o outro com os dados estruturados. Os resultados são mostrados na Figura 6.6 e na Tabela 6.9.

Com a utilização de TF-IDF e dados estruturados em duas redes em paralelo, o resultado balanceado de F1-*score* do melhor modelo foi de $90,67\%$ para a fusão de diferença

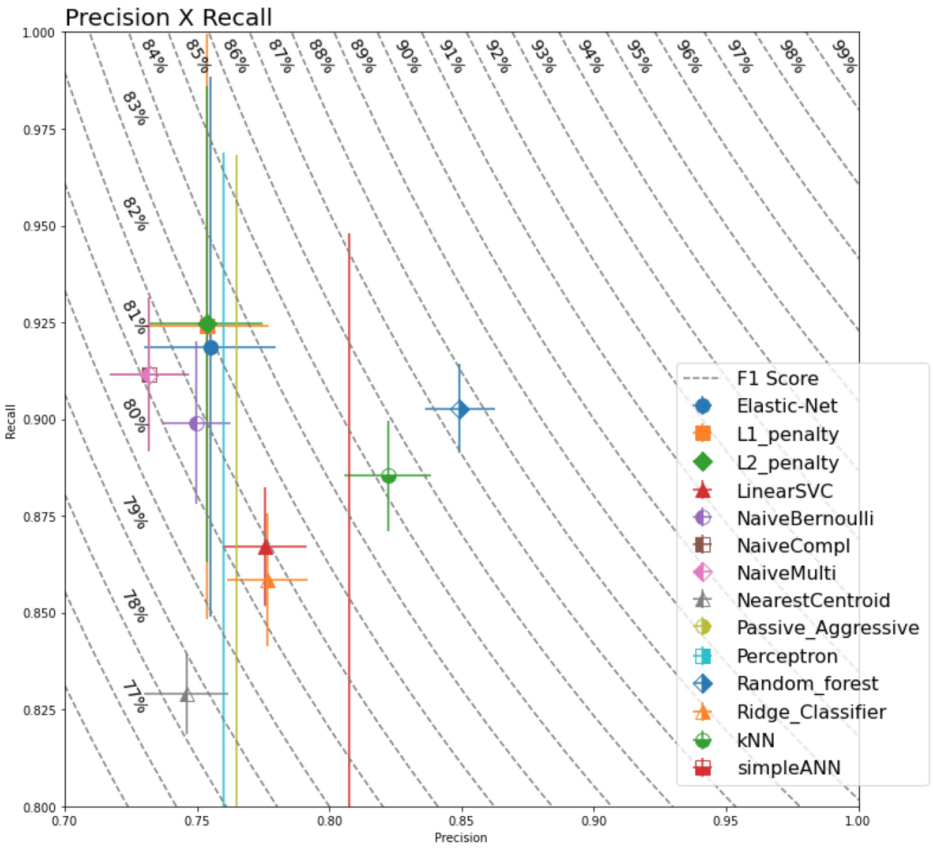


Figura 6.5: Precisão, sensibilidade e F1-*score* dos modelos lineares dos dados estruturados.

Tabela 6.9: Valores encontrados na validação cruzada dos modelos de redes multicanal *bottleneck* com TF-IDF e dados estruturados.

Método	**Média**		
	Precisão	**Sensibilidade**	**F1-Score**
BN-Str-abd	93.30%±3.85%	87.48%±4.88%	90.20%±3.53%
BN-Str-add	93.49%±2.42%	**87.74%±3.95%**	90.43%±1.47%
BN-Str-avg	93.30%±3.00%	86.09%±4.79%	89.41%±2.23%
BN-Str-con	93.60%±2.14%	87.05%±4.07%	90.11%±1.58%
BN-Str-mul	**94.55%±1.78%**	84.70%±3.60%	89.29%±1.72%
BN-Str-sqd	93.54%±3.43%	88.26%±4.68%	**90.67%±1.78%**
BN-Str-sub	94.08%±1.41%	86.68%±2.44%	90.20%±1.25%

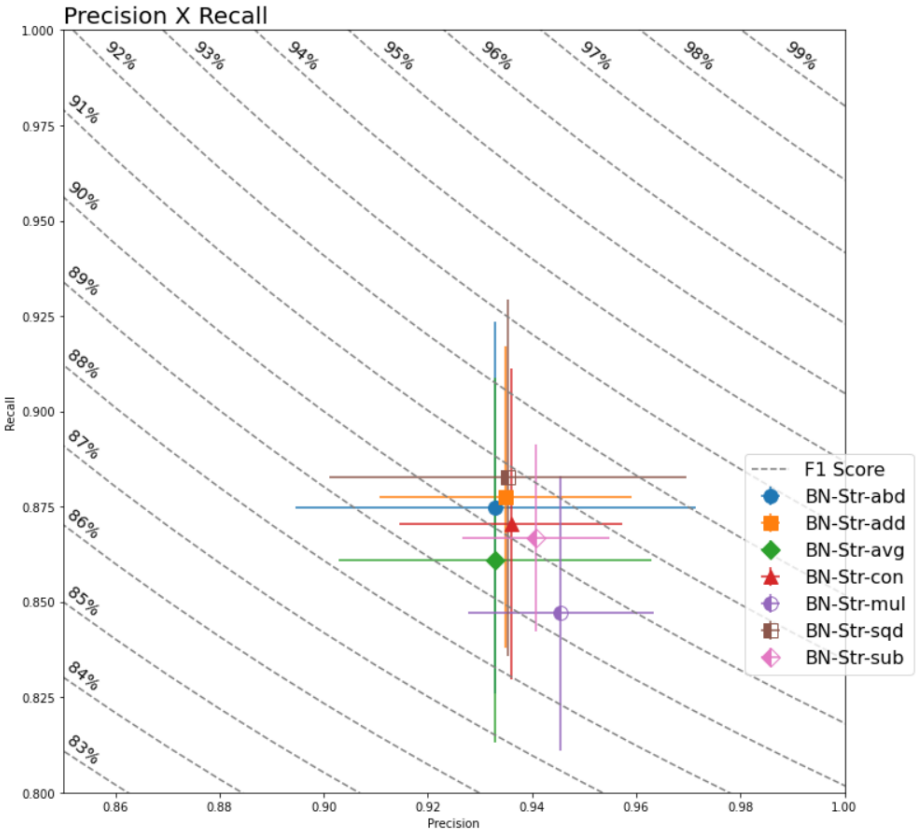


Figura 6.6: Precisão, sensibilidade e F1-*score* dos modelos multicanal *bottleneck* com TF-IDF e dados estruturados.

quadrada (sqd), sendo este pior que os resultados anteriores. Entretanto, ocorreu uma melhora significativa na precisão para o modelo que utilizou multiplicação na fusão chegando a $94,55\%$, este melhor que o valor de $92,8\%$ observado para a rede de canal único *bottleneck* com TF-IDF. Infelizmente, a sensibilidade ficou baixa nestes casos, variando entre $84,7\%$ (multiplicação) e $87,74\%$ (adição).

Ao comparar os melhores resultados, mostrando todos os pontos da validação cruzada e os limites de precisão e sensibilidade (casco convexo dos pontos) na Figura 6.7, fica evidente que os casos de modelos multicanais apresentam uma maior variabilidade de sensibilidade (*recall*) e precisão, com variações de mais de 20%. Assim, os melhores modelos ficam entre o *passive-aggressive* - para uma escolha de melhor F1-*score* - e o *bottleneck* - para uma escolha de melhor precisão.

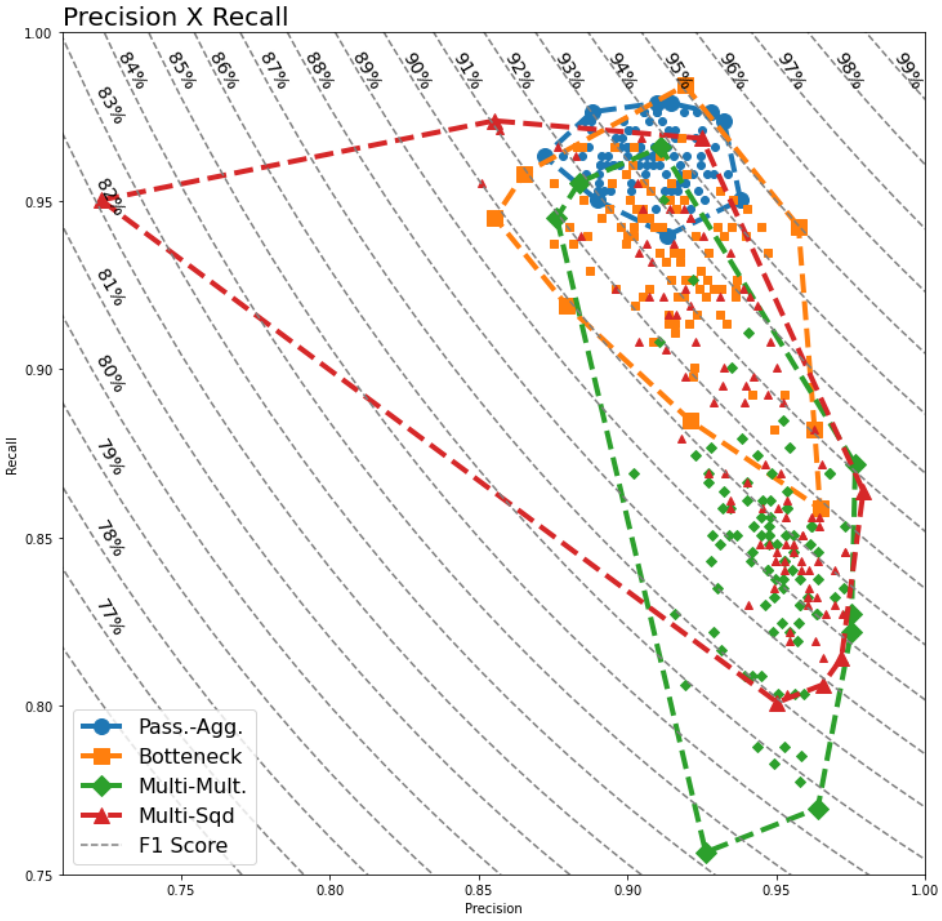


Figura 6.7: Precisão, sensibilidade e F1-*score* dos melhores modelos os limites de precisão e sensibilidade (casco convexo dos pontos).

6.5 Limiar e *Equal Error Rate* - EER (Melhores Modelos)

Para esta seção foi selecionado dentro dos conjuntos de validação cruzada os modelos que geraram melhores resultado de F1-*score* para o modelo linear *Passive-Aggressive*, para que assim houvesse o maior potencial de bons resultados em uma predição de risco no futuro.

Seguindo os passos da Seção 2.5, foi feito o cálculo do limiar (EER) de decisão para o modelo *Bottleneck* que apresentou resultados melhores do ponto de vista da precisão. Esta análise se faz necessária para definir o ponto de decisão do limiar ao qual, em termos de classificação, pertencerá a saída do classificador. Esta informação tem como base a avaliação estatisticamente coerente (ou com significância) no uso do classificador.

O modelo linear já apresenta equilíbrio no seu limiar, ou seja, o limiar calculado pelo EER já é de 50%.

O gráfico da Figura 6.8, baseado na taxa de falsos positivos (FPR) e na taxa de falsos negativos (FNR), mostra o ponto de encontro das duas curvas no limiar de $0,999$. Com o limiar encontrado as publicações que tiverem a saída do modelo abaixo de $0,999$ devem ser classificados como *risco 0* e acima deste valor deve ser classificado como *risco 1*.

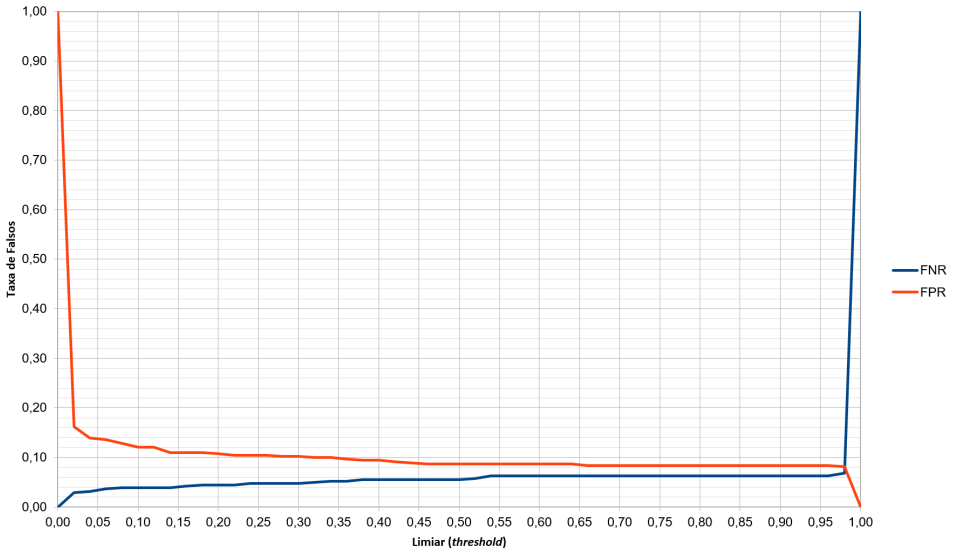


Figura 6.8: Cálculo do limiar decisório baseado na taxa de falsos positivos (FPR) e na taxa de falsos negativos (FNR).

De forma equivalente o limiar também pode ser obtido pela curva ROC (Figura 6.9), por meio do cruzamento desta curva com a diagonal da função $f(x) = -x + 1$ e, como esperado, se obtém o mesmo valor de $0,999$.

A partir dos mesmos dados e gráfico da Figura 6.9 se obtém o valor de áreas sob a ROC (AUC) de 97,92% para o modelo *Bottleneck* e 97,33% para o modelo *Passive-Aggressive*. Estes valores estão muito próximos e entrariam em uma classificação de "ótimo" segundo a Figura 2.11[59].

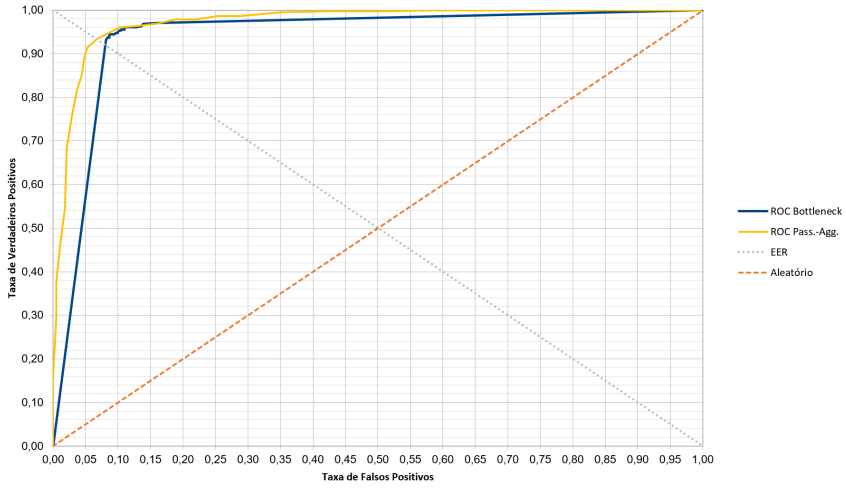


Figura 6.9: Curva ROC para os modelos *Bottleneck* e *Passive-Aggressive*).

A Figura 6.10 mostra as variações da matriz de confusão antes e depois de fazer a classificação com ajuste pelo limiar para o modelo *Bottleneck* onde observa-se que, neste caso, há diminuição dos Verdadeiros Positivos com aumento dos Verdadeiro Negativos para se chegar a um equilíbrio. Para o modelo *Passive-Aggressive*, na mesma figura, observa-se que este modelo só não foi melhor nas taxas de Verdadeiros Negativos e Falsos Positivos.

Já a Tabela 6.10 mostra como a alteração do limiar faz com que haja conversão da precisão e da sensibilidade para um mesmo valor para o modelo *Bottleneck* e como o modelo *Passive-Aggressive* apresentou melhores resultados em toda as métricas.

Tabela 6.10: Precisão, Sensibilidade e F1-*score* com e sem a alteração do limiar em comparação com melhor conjunto de treinamento do modelo *Passive-Agressive*.

Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-*score***
Bottleneck Limiar=0,5	92,54%	91,6%	94,5%	93,0%
Bottleneck Limiar=0,999	92,68%	92,7%	92,7%	92,7%
Passive-Aggressive	92,54%	93,0%	97,4%	95,1%

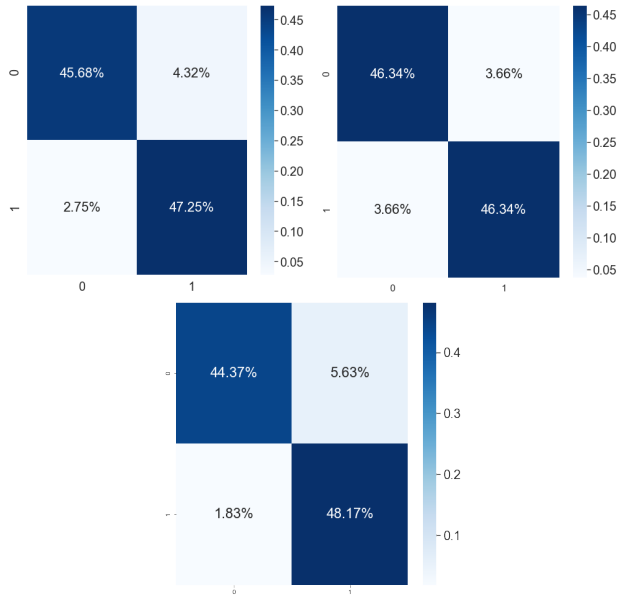


Figura 6.10: Comparação das Matrizes de Confusão sem (a) e com (b) alteração do limiar para o modelo *Bottleneck* e (c) para o modelo *Passive-Aggressive*.

7

Conclusão e Trabalhos Futuros

Somos fortes, na linha avançada,
sem da luta os embates temer,
que à chamada da Pátria insultada,
saberemos cumprir com o dever.

– Hino da Polícia Federal - Eugênio Lapagesse

7.1 Conclusões

Este trabalho apresentou a formação de um banco de dados anotado para classificação de texto na detecção de risco de fraude em obras de engenharia. Bem como, fora realizada a comparação de redes neurais profundas, *autoencoders*, Bi-LSTM com classificadores lineares clássicos, na classificação do risco de conluio em obras, a partir de documentos produzidos e publicados por fontes públicas de informações.

O uso de classificadores clássicos e de redes neurais artificiais profundas provou que técnicas de aprendizagem de máquina e processamento de linguagem natural são capazes de classificar o conjunto de dados de publicações e chegar a um modelo confiável para classificação de risco. Os modelos foram avaliados quanto ao conjunto de características: precisão, sensibilidade e de F1-*score*, para fins quantitativos.

Entre os modelos de extração de características, o TF-IDF mostrou-se o melhor desempenho para o conjunto de dados e os melhores classificadores de redes clássicas e neurais obtiveram resultados de F1-*score* acima de 93,0%. Os modelos de redes neurais profundas *Bottleneck* e Bi-LSTM, provaram serem competitivos com os classificadores tradicionais e obtiveram boa precisão, o que é mais desejável em uma investigação de fraude criminal. Entretanto modelo *Passive-Aggressive* apresentou melhores resultados em toda as métricas com F1-*score* de 95,1% para o melhor conjunto de validação cruzada.

Desta forma, foi alcançado o objetivo principal de se obter um classificador de risco das publicações do Diário Oficial da União como forma de detectar indícios de fraudes e conluios em licitações de obras públicas no Brasil. Além disso, os objetivos secundários também foram alcançados: o banco de dados anotados foi formado, foram identificadas publicações vinculadas a risco, foi constatado que o TF-IDF é uma forma de representação adequada à classificação destes tipos de textos e foram comparados classificadores de textos do estado da arte em todo o processo realizado.

Também como resultado dos trabalhos realizados, foi publicado o artigo científico no *Findings of the Association for Computational Linguistics: EMNLP 2020* (**Qualis/CAPES CC A1**) com o título *Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach*[85] (texto completo publicado em Apêndice, página 103 e seguintes).

7.2 Trabalhos Futuros

Vislumbra-se a possibilidade de desenvolvimento de técnicas para melhorar e customizar a extração de recursos para o conjunto de dados específico. Como a utilização de dicionários e *embeddings* criados e desenvolvidos para publicações do Diário Oficial.

Dada a ampla gama de customização de redes neurais, espera-se, por exemplo, que as técnicas de *data augmentation* melhorem os resultados, estes devidos à pequena quantidade de dados anotados disponíveis no conjunto de dados proposto. Ainda assim, é uma das muitas maneiras de obter melhorias de desempenho no processo de classificação. Uma abordagem, por exemplo usando redes adversarias generativas, na linha proposta em [71, 72], permitiria a criação destes novos dados artificiais, reduzindo as dificuldades encontradas na formação do conjunto de dados.

Pode ganhar maior confiabilidade do processo classificatório se com a utilização do Sistema *Deep Vacuity* houver retroalimentação com a crescente base de publicações do tipo *risco 1*, como novos casos advindos das investigações em curso, bem como permitindo a inserção de uma quantidade de informações relevante com o uso desta ferramenta dentro do contexto policial de atividades.

A base de dados construída, agora em sua totalidade, ainda pode ser utilizada para gerar outras fontes de dados que permitam a `comprovação` de outras ilicitudes, como por exemplo cartéis, utilizando as técnicas desenvolvidas por Signor et al. [115], dentre outras desenvolvidas e apresentadas na literatura atualmente disponível, do ponto de vista de análise de contratos e obras públicas.

Referências

[1] Andre Abade., Ana S. de Almeida., and Flavio Vidal. Plant diseases recognition from digital images using multichannel convolutional neural networks. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 450–458. INSTICC, SciTePress, 2019. ISBN 978-989-758-354-4. doi: 10.5220/0007383904500458. 33

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org. 65

[3] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015. 24

[4] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. 65

[5] Jose María Alvarez, José Emilio Labra, Ángel Marín, and José Luis Marín. Semantic methods for reusing linking open data of the european public procurement notices. In *ESWC PhD Symposium*, 2011. 9, 47

[6] Jose María Alvarez-Rodríguez, Michalis Vafopoulos, and Juan Llorens. Enabling policy making processes by unifying and reconciling corporate names in public procurement data. the corfu technique. *Computer Standards & Interfaces*, 41:28 – 38, 2015. ISSN 0920-5489. doi: https://doi.org/10.1016/j.csi.2015.02.009. URL http://www.sciencedirect.com/science/article/pii/S0920548915000252. 9

[7] Hubert Anysz, Andrzej Foremny, and Janusz Kulejewski. Comparison of ann classifier to the neuro-fuzzy system for collusion detection in the tender procedures of road construction sector. In *IOP Conference Series: Materials Science and Engineering*, volume 471, page 112064. IOP Publishing, 2019. 9, 46

[8] APCF. Perícia criminal. http://apcf.org.br/pericia-criminal/pericia-criminal, 2020. visited: 2020-05-25. 7

[9] Remis Balaniuk, Pierre Bessiere, Emmanuel Mazer, and Paulo Cobbe. Risk based Government Audit Planning using Naïve Bayes Classifiers. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, Spain, 2012. doi: 10.3233/978-1-61499-105-2-1313. URL https://hal.archives-ouvertes.fr/hal-00746198. 45

[10] Rebeca Andrade Baldomir. *Aplicação do Algoritmo Apriori para Detectar Relacionamentos entre Empresas nos Processos Licitatórios do Governo Federal*. PhD thesis, Universidade de Brasília, 2017. 9

[11] András A. Benczúr, Levente Kocsis, and Róbert Pálovics. Online machine learning in big data streams, 2018. 64

[12] Lucio Big. Operação política supervisionada. https://www.ops.net.br/, oct 2018. 41

[13] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 01 2009. ISBN 978-0-596-51649-9. 21, 22

[14] Leon Bottou. Stochastic gradient descent. https://leon.bottou.org/projects/sgd, 2010. Visited: 2020-05-24. 64

[15] Max Bramer. *Principles of data mining*, volume 180. Springer, 2007. xi, 19, 35

[16] Brasil. Decreto-lei 73.332, de 19 de dezembro de 1973. http://www.planalto.gov.br/ccivil_03/decreto/antigos/d73332.htm, 1973. Acessado: 2021-03-26. 6

[17] Brasil. Lei n. 8.666, de 21 de junho de 1993. http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm, 1993. Accessed: 2019-11-06. xi, 9, 12, 13, 15, 16, 18, 55

[18] Brasil. Lei n. 10.520, de 17 de julho de 2002. http://www.planalto.gov.br/ccivil_03/LEIS/2002/L10520.htm, 2002. Accessed: 2019-11-06. xi, 9, 12, 13, 15, 18, 55

[19] Brasil. Lei 10.683, de 28 de maio de 2003. http://www.planalto.gov.br/ccivil_03/leis/2003/l10.683.htm, 2003. Acessado: 2021-03-26. 6

[20] Brasil. Decreto 6.017, de 17 de janeiro de 2007. https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6017.htm, 2007. Acessado: 2021-03-26. 14

[21] Brasil. Decreto 6.170, de 25 de julho de 2007. https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6170.htm, 2007. Acessado: 2021-03-26. 13

[22] Brasil. Lei 14.133, de 4 de agosto de 2011. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12462.htm, 2011. Acessado: 2021-04-09. 12, 15, 18

[23] Brasil. Decreto 9.215, de 29 de novembro de 2017. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/D9215.htm, 2017. Acessado: 2021-04-09. 10

[24] Brasil. Constituição da república federativa do brasil de 1988. http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm, 2020. Acessado: 2021-03-26. 6

[25] Brasil. Lei 14.133, de 1º de abril de 2021. http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm, 2021. Acessado: 2021-04-09. 7, 18

[26] Ministério da Economia Brasil. 020307 - transferÊncias voluntÁrias - manual sadipem. https://conteudo.tesouro.gov.br/manuais/index.php?option=com_content&view=article&id=1543:020307-transferencias-voluntarias&catid=749&Itemid=376, 2020. Visited: 2020-11-13. 13

[27] Open Knowledge Brasil. Querido diário. https://github.com/okfn-brasil/querido-diario, 2020. 54

[28] Fabricio Ataides Braz, Nilton Correia da Silva, Teofilo Emidio de Campos, Felipe Borges S Chaves, Marcelo HS Ferreira, Pedro Henrique Inazawa, Victor HD Coelho, Bernardo Pablo Sukiennik, Ana Paula Goncalves Soares de Almeida, Flavio Barros Vidal, et al. Document classification using a bi-lstm to unclog brazil's supreme court. *arXiv preprint arXiv:1811.11569*, 2018. 49

[29] Brazil. Imprensa nacional. http://www.in.gov.br, 2020. visited: 2020-04-23. xi, xiv, 10, 15, 16, 17, 55, 57

[30] Brazil. Painel de obras. http://transferenciasabertas.planejamento.gov.br/, 2020. visited: 2020-05-19. 7

[31] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 65

[32] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992. 23, 61

[33] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 51, 52

[34] Amelie Byun, Fei-Fei Li, Ranjay Krishna, and Dnafei Xu. Cs231n: Convolutional neural networks for visual recognition. https://cs231n.github.io/, 2010. Visited: 2020-09-24. 26

[35] CADE. Combate a cartéis em licitações. guia prático para pregoeiros e membros de comissões de licitação, 2008. 6, 8

[36] Jose Camacho-Collados and Mohammad Taher Pilehvar. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*, 2017. 20

[37] Jose Camacho-Collados and Mohammad Taher Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, 2018. 24

[38] Ricardo Silva Carvalho and Rommel Novaes Carvalho. Bayesian models to assess risk of corruption of federal management units. In *BMA@ UAI*, pages 28–35, 2016. 46

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, Jun 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953. 46

[40] Junyi Chen, Shankai Yan, and Ka-Chun Wong. Verbal aggression detection on twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, pages 1–10, 2018. 49

[41] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230, 2017. 49

[42] Deysi Cioccari. Operação lava jato: escândalo, agendamento e enquadramento. *Revista Alterjor*, 12(2):58–78, out. 2015. URL https://www.revistas.usp.br/alterjor/article/view/aj12-a04. 40

[43] APIS Company. Biometric principles. https://biometria.apis.sk/en/principles-of-biometrics.html, 2020. Visited: 2020-11-23. xii, 38

[44] Marina Cordeiro et al. Acerca da (des) necessidade do dolo específico e do prejuízo ao erário para a configuração dos crimes previstos nos arts. 89 e 90 da lei de licitações, 2016. 8

[45] R Gallisson D COSTE and Robert GALLISSON. Dicionário de didáctica das línguas. *Coimbra, Editora Almedina*, 1983. 22

[46] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006. 64, 65

[47] Controladoria Geral da União. Observatório da despesa pública, oct 2018. URL http://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa-publica. 41

[48] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Language modeling with longer-term dependency. 2018. 52

[49] A.P.G.S. de Almeida and F. de Barros Vidal. L-cnn: a lattice cross-fusion strategy for multistream convolutional neural networks. *Electronics Letters*, 55 (22):1180–1182, 2019. doi: https://doi.org/10.1049/el.2019.2631. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/el.2019.2631. 33

[50] Pedro Henrique Luz De Araujo. From Documents to Entities: A journey through Natural Language Processing tasks and domains. Master's thesis, Universidade Federal de Brasília, Brasília, 2020. 9

[51] Pedro Henrique Luz De Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. Victor: a dataset for brazilian legal documents classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1449–1458, 2020. 51

[52] Pedro Henrique Luz De Araujo, Teófilo Emidio de Campos, and Marcelo Magalhães Silva de Sousa. Inferring the source of official texts: can svm beat ulmfit? In *International Conference on Computational Processing of the Portuguese Language*, pages 76–86. Springer, 2020. 25, 52

[53] Talita Lôbo de Menezes. Eficácia no uso de aprendizagem de máquina para estimação de risco em contratos públicos e empresas. Master's thesis, Centro de Engenharia Elétrica e Informática, 2019. Coordenação de Pós-Graduação em Ciência da Computação. 46

[54] V. de Oliveira Silva, F. de Barros Vidal, and A. R. Soares Romariz. Human action recognition based on a two-stream convolutional network classifier. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 774–778, 2017. doi: 10.1109/ICMLA.2017.00-64. 33

[55] Xuelian Deng, Yuqing Li, Jian Weng, and Jilian Zhang. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3):3797–3816, 2019. 49

[56] Matthew Denny and Arthur Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *When It Misleads, and What to Do about It (September 27, 2017)*, 2017. 20, 61

[57] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 51, 52

[58] Andre Luiz dos Santos Nakamura. A infraestrutura e a corrupção no brasil. *Revista Brasileira de Estudos Políticos*, 117, 2018. 7

[59] Rachel Lea Ballantyne Draelos. Measuring performance: Auc (auroc). https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/, 2019. Visited: 2020-10-31. xii, 37, 88

[60] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008. 65

[61] Caixa Econômica FEDERAL. Sinapi: Metodologias e conceitos, 2017. 8

[62] Supremo Tribunal Federal. Ministra cármen lúcia anuncia início de funcionamento do projeto victor, de inteligência artificial. http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=388443, 2020. Visited: 2020-11-23. 49

[63] Shuo Feng, Huiyu Zhou, and Hongbiao Dong. Using deep neural network with small dataset to predict material defects. *Materials & Design*, 162:300–310, 2019. 64

[64] Arthur Emidio T. Ferreira., Ana Paula G. S. de Almeida., and Flavio de Barros Vidal. Autonomous vehicle steering wheel estimation from a video using multichannel convolutional neural networks. In *Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO,*, pages 517–524. INSTICC, SciTePress, 2018. ISBN 978-989-758-321-6. doi: 10.5220/0006920605170524. 33

[65] Hugo Honda Ferreira. Processamento de Linguagem Natural e Classificação de Textos em Sistemas Modulares., 2018. Monografia (Bacharel em Informática), Universidade de Brasília, Brasília, Brazil. 54, 74

[66] José Roberto Ferro. O departamento de "operações estruturadas" da odebrecht. https://epocanegocios.globo.com/colunas/Enxuga-Ai/noticia/2016/03/o-departamento-de-operacoes-estruturadas-da-odebrecht.html, mar 2016. 41

[67] Open Knowledge Foundation. Procurement. http://index.okfn.org/dataset/procurement/, 2020. visited: 2020-05-19. 9

[68] Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999. 65

[69] Sunita Goel and Ozlem Uzuner. Do sentiments matter in fraud detection? estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239, 2016. 48, 63

[70] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017. xi, 18, 19, 23, 25, 26, 27, 28, 31, 34

[71] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 28, 91

[72] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 91

[73] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE, 2005. 33, 66

[74] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017. 65

[75] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 31

[76] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 52

[77] Subbu Kannan and Vairaprakash Gurusamy. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5 (1):7–16, 2014. 22

[78] Kei Kawai and Jun Nakabayashi. Detecting large-scale collusion in procurement auctions. 2018. 9

[79] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. 30

[80] Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327, 2016. 49

[81] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019. xi, 19, 20, 49

[82] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991. 30

[83] Lanker Vinnícius Borges Silva Landin. A impunidade e a selectividade dos crimes de colarinho branco. Master's thesis, Direito, Relações Internacionais e Desenvolvimento, 2015. Ciências Humanas. 40

[84] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014. 62, 65

[85] Marcos Lima, Roberta Silva, Felipe Lopes de Souza Mendes, Leonardo R. de Carvalho, Aleteia Araujo, and Flavio de Barros Vidal. Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1580–1588, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.143. URL https://www.aclweb.org/anthology/2020.findings-emnlp.143. 91, 103

[86] Marcos Cavalcanti Lima. Comparação de custos referenciais do dnit e licitações bem sucedidas. *Revista do Tribunal de Contas da União (TCU) nº*, 110:59, 2010. 8

[87] Gang Liu and Jiabao Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.01.078. URL https://www.sciencedirect.com/science/article/pii/S0925231219301067. 49

[88] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 49, 52

[89] Alan de Oliveira Lopes. Superfaturamento de obras públicas. *Livro Pronto, São Paulo, Brazil*, 2011. 7

[90] Alan de Oliveira Lopes. O efeito pedagógico de operações da polícia federal: Um estudo de caso da operação"caixa de pandora". *Revista Brasileira de Ciências Policiais*, 6(1):67–85, 2015. 8

[91] Hans Peter Luhn. Key word-in-context index for technical literature (kwic index). *American Documentation*, 11(4):288–295, 1960. 22

[92] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999. 24

[93] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008. 24, 35, 65

[94] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. 24

[95] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*, 2020. xii, 50

[96] Irio Musskopf. Operação serenata de amor. https://serenata.ai/about/, oct 2018. 41

[97] OCDE. Bribery in public procurement methods, actors and counter-measures. 2007. 7

[98] Christopher Olah. Understanding lstm networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2015. Visited: 2020-10-21. xi, 32, 33

[99] Bruno Oliveira. O duplo grau de jurisdição na ação penal 470/mg: Considerações à luz do controle de convencionalidade. *Revista Direito em Debate*, 26(47):267–288, set. 2017. doi: 10.21527/2176-6622.2017.47.267-288. URL https://www.revistas.unijui.edu.br/index.php/revistadireitoemdebate/article/view/5771. 41

[100] V Pareto. Cours d'économie politique, nouvelle édition par gh bousquet et g. *Busino, Genève, Droz*, 1, 1964. 34

[101] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A rule based approach to word lemmatization. In *Proceedings of IS*, volume 3, pages 83–86, 2004. 23

[102] Fred Popowich. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1):59–66, 2005. 48, 63

[103] Peter Prettenhofer, Olivier Grisel, Mathieu Blondel, and Lars Buitinck. Classification of text documents using sparse features. https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html, 2019. Accessed: 2020-04-25. xiv, 64, 65

[104] Reza Rajabiun and Catherine Middleton. Public interest in the regulation of competition: Evidence from wholesale internet access consultations in canada. *Journal of Information Policy*, 5:32–66, 2015. 47

[105] Célia Ghedini Ralha and Carlos Vinícius Sarmento Silva. A multi-agent data mining system for cartel detection in brazilian government procurement. *Expert Systems with Applications*, 39(14):11642 – 11656, 2012. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2012.04.037. URL http://www.sciencedirect.com/science/article/pii/S0957417412006343. 9, 45

[106] Fuji Ren and Jiawen Deng. Background knowledge based multi-stream neural network for text classification. *Applied Sciences*, 8(12):2472, Dez 2018. ISSN 2076-3417. doi: 10.3390/app8122472. URL http://dx.doi.org/10.3390/app8122472. 33

[107] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003. 65

[108] Ryan M Rifkin and Ross A Lippert. Notes on regularized least squares. 2007. 65

[109] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 51

[110] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 66

[111] Prasad Seemakurthi, Shuhao Zhang, and Yibing Qi. Detection of fraudulent financial reports with machine learning techniques. In *2015 Systems and Information Engineering Design Symposium*, pages 358–361. IEEE, 2015. 48, 63

[112] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Bi-directional block self-attention for fast and memory-efficient sequence modeling. *arXiv preprint arXiv:1804.00857*, 2018. 49, 51

[113] Regis Signor, Peter ED Love, João José CB Vallim, Alexandre B Raupp, and Oluwole Olatunji. It is not collusion unless you get caught: the case of operation car wash and unearthing of a cartel. *Journal of Antitrust Enforcement*, 2019. 8, 9, 45

[114] Regis Signor, Peter ED Love, Alexanders TN Belarmino, and Oluwole Alfred Olatunji. Detection of collusive tenders in infrastructure projects: Learning from operation car wash. *Journal of Construction Engineering and Management*, 146(1): 05019015, 2020. 45

[115] Regis Signor, Peter ED Love, Acir Oliveira Jr, Alan O Lopes, and Pedro S Oliveira Jr. Public infrastructure procurement: Detecting collusion in capped first-priced auctions. *Journal of Infrastructure Systems*, 26(2):05020002, 2020. 45, 91

[116] Carlos Vinícius Sarmento Silva. *Agentes de Mineração e sua Aplicação no Domínio de Auditoria Governamental*. PhD thesis, Universidade de Brasília, 2011. 9

[117] Laércio de Oliveira Silva Filho, Marcos Cavalcanti Lima, and Rafael Gonçalves Maciel. Efeito barganha e cotação: fenômenos que permitem a ocorrência de superfaturamento com preços inferiores às referências oficiais. *Revista do Tribunal de Contas da União. Brasília*, pages 29–36, 2010. 7

[118] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. xii, 29, 76, 77

[119] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018. xi, 29, 30, 76

[120] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 29

[121] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019. 49, 51, 52

[122] Ting Sun and Leonardo J. Sales. Predicting Public Procurement Irregularity: An Application of Neural Networks. *Journal of Emerging Technologies in Accounting*, 15(1):141–154, 03 2018. ISSN 1554-1908. doi: 10.2308/jeta-52086. URL https://doi.org/10.2308/jeta-52086. 9, 45

[123] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002. 65

[124] DE CONTAS DA UNIÃ O TRIBUNAL. *Recomendações Básicas para a Contratação e Fiscalização de Obras de Edificações Públicas*. Tribunal de Contas da União, 2014. 15, 16

[125] João José de Castro Baptista Vallim. Uso do Modelo de Raciocínio Baseado em Casos Para Monitoramento de Conluio em Licitações de Obras de Pavimentação Urbana. Master's thesis, Universidade Federal do Paraná, Curitiba, 2020. 9, 44

[126] John H Van Arkel, James J Wagner, Corrine L Schweyen, Saralyn M Mahone, Terrill J Curtis, Scott HAGINS, et al. Predictive modeling processes for healthcare fraud detection, January 3 2013. US Patent App. 13/536,414. 48, 63

[127] S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015. 20

[128] Bo Pedersen Weidema and Marianne Suhr Wesnaes. Data quality management for life cycle inventories—an example of using data quality indicators. *Journal of cleaner production*, 4(3-4):167–174, 1996. 55

[129] Thomas Wood. What is the softmax function? https://deepai.org/machine-learning-glossary-and-terms/softmax-layer, 2010. Visited: 2020-09-24. 27

[130] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 52

[131] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002. 65

[132] Tong Zhang, Fred Damerau, and David Johnson. Text chunking based on a generalization of winnow. *Journal of Machine Learning Research*, 2(Mar):615–637, 2002. 65

Apêndice

Artigo publicado no *Findings of the Association for Computational Linguistics: EMNLP 2020* com o título *Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach*[85].

Disponível em: https://www.aclweb.org/anthology/2020.findings-emnlp.143/

Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach

Marcos Cavalcanti Lima
Dept. of Computer Science
University of Brasília (UnB)
Brasília - DF - Brazil
marcosc.mcl@pf.gov.br

Leonardo R. de Carvalho
Dept. of Computer Science
University of Brasília (UnB)
Brasília - DF - Brazil
leouesb@gmail.com

Felipe L. de Souza Mendes
Dept. of Computer Science
University of Brasília (UnB)
Brasília - DF - Brazil
felipelopess@gmail.com

Roberta da Costa Silva
Dept. of Computer Science
University of Brasília (UnB)
Brasília - DF - Brazil
robertacs42@gmail.com

Aleteia Araujo
Dept. of Computer Science
University of Brasília (UnB)
Brasília - DF - Brazil
aleteia@unb.br

Flávio de Barros Vidal
Dept. of Computer Science
University of Brasília (UnB)
Brasília - DF - Brazil
fbvidal@unb.br

Abstract

Public works procurements move US$ 10 billion yearly in Brazil and are a preferred field for collusion and fraud. Federal Police and audit agencies investigate collusion (bid-rigging), over-pricing, and delivery fraud in this field and efforts have been employed to early detect fraud and collusion on public works procurements. The current automatic methods of fraud detection use structured data to classification and usually do not involve annotated data. The use of NLP for this kind of application is rare. Our work introduces a new dataset formed by public procurement calls available on Brazilian official journal (*Diário Oficial da União*), using by 15,132,968 textual entries of which 1,907 are annotated risky entries. Both bottleneck deep neural network and Bi-LSTM shown competitive compared with classical classifiers and achieved better precision (93.0% and 92.4%, respectively), which signs improvements in a criminal fraud investigation.

1 Introduction

In the last five years, Brazil's federal government invested (Brazil, 2020b) in 23,352 public works contracts adding up to R$ 283.8 billion (approx. US$ 49.3 billion in May 2020). Those works consist of all sorts of projects from oil refineries to ports, from soccer stadiums to power plants, from tunnels to dams, and are developed on a continental-sized territory, generating an endless and growing quantity of information regarding those projects. Thereupon, public works procurements are a preferred field for collusion and fraud (OCDE, 2007).

Brazilian Federal Police have been working on fraud investigations on public works for the last four decades and develop its investigation based on a highly skilled group of experts formed by civil, electrical, mechanical, computer engineers, and accountants (APCF, 2020). The types of fraud investigated are mainly collusion (bid-rigging), over-pricing, and delivery fraud (quality and quantity of services and materials). We will bring the knowledge accrued during those decades to enhance our data understanding (Lopes, 2011).

As described in Foundation (2020), public works contracts are made via procurement, that is the process of public administration uses to make all its contracts. Every procurement step is usually publicized by a call for application, and any interested people (or enterprises) around the world can obtain data from all available government journals of Brazil (the prominent public information journal in Brazil is the *Diário Oficial da União - DOU*). Despite being easy to access, tables, texts and documents do not bring any other annotated data for classification, even less for fraud detection. Those types of datasets have been studied, for example, by named-entity recognition in (Alvarez-Rodríguez et al., 2015) and linking open data, as in (Alvarez et al., 2011).

On the other hand, detecting and proving fraud on construction procurements is a laborious task, consuming around one month of forensic expert work per procurement/contract. Furthermore, it is essential to detect and combat fraud since a procurement first step because, as observed (Signor et al., 2019; Lopes, 2015), over-pricing is hardly obtained without collusion as most prices are set

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1580–1588

during procurement. So, it has been the object of many studies (Kawai and Nakabayashi, 2014; Signor et al., 2019; Anysz et al., 2019; Sun and Sales, 2018; Vallim, 2020), but none of them used unstructured data to produce its goals.

Based on these presented statements, this work is focused on present a new dataset with textual information extracted from Brazilian Public government journals with an annotated *ground-truth* by forensics experts. It is, also, included in this work an initial classification methodology using a Bi-LSTM model and all early results are compared with main *state of the art* techniques.

The manuscript is organized as follows: In Section 2 are presented the related works about fraudulent collusion on public works contracts in official texts. Section 3 explains our proposed methodology. Section 4 contains information about the results and discussion. Section 5 provides conclusion points and introduce further works.

2 Related Works

It will be presented two parts: the first one, about fraud detection efforts, and the second one, about NLP classification.

Brazilian Federal Police has aspiring to improve its fraud detection mechanisms. Therefore, as presented in Vallim (2020) made a CBR model of paving services in the Paraná State approaching paving works contracts, which are one of the most budget consuming services in a state or city level and focus of criminal activities. This model used procurement type, enterprises, contract, and georeferenced data, and aimed to classify collusion cases, all of them based in a manual approach.

Another way to prove and identify public procurement collusion is by the use of statistics and probability. Those methods were explored on several Federal Police's studies and were based on joint behavior analysis of competitors who act to achieve bid-rigging. It was successfully applied to oil-related contracts using Operation Car Wash[1] information (Signor et al., 2019, 2020a) and for infrastructure projects (Signor et al., 2020b) with capped first-price auctions.

Brazilian Comptroller General of the Union (CGU), a national auditing public agency, also has several initiatives to reach a reliable classifier for public procurement fraud. In Ralha and Silva (2012) was elaborated a unsupervised evaluator that, using *a priori* rules, evaluate the possibility of a certain group winning a given tender. They used structured data to bring suspect groups to be evaluated by experts. The work developed in Balaniuk et al. (2012) focused on the evaluation of fraud risk in government agencies using a Naïve Bayes Classifier for audit planning by the use of structured data and patterns of fraudulent activity. Sun and Sales (2018) used traditional neural networks and deep neural networks (DNN) to elaborate an early alarm system. The CGU studies usually have as features and fraud indicators: the number of bids, estimated cost and price relations, relations between public and private parts, political links of political parties, etc. Carvalho and Carvalho (2016) achieved good results using Bayesian Models with structured data from penalties database. They used data enrolled from the federal civil servants, servants' roles and income, number of accounts judged irregular and number of regularity certificates on an agency unit, and affiliated servants of each management unit.

Anysz et al. (2019) uses ANN and structured data on Poland's highways public procurements. They used the number of enterprises, price differences, contract order in the same place, and set of propositions to assess its fraud risk.

Works using TF-IDF in procurement documentation, as presented by Rabuzin and Modrušan (2019), tested Logistic Regression, SVM and Naïve Bayes on potential corruption. Their model had no annotated data, so it was focused in finding one bid tenders which "could be potentially suspicious."

Natural Language Processing is not often used to classify public procurement documents for risk or fraud. The technology is used for assessing fraud risk in health care claims (Popowich, 2005; Van Arkel et al., 2013) and financial reports (Seemakurthi et al., 2015; Goel and Uzuner, 2016). Public works publications data are not uniform enough to be structured, and, even if it is possible, it would be extremely laborious and it might be done at the cost of losing some unknown or undetected features.

All these studies suggest that fraud or risk public procurement classification has been developed based mostly on structured data and the use of NLP for this specific kind of classification is rare or nonexistent.

Regarding NLP classification methods, Braz

[1]Operation Car Wash is an ongoing nationwide corruption investigation led by Brazilian Federal Police, and it is focused on Petrobrás procurements. It is called "the biggest corruption scandal in history" (Watts, 2017).

et al. (2018) proposed a Bi-LSTM based model to classify Brazilian supreme court documents as part of VICTOR project. de Araujo et al. (2020) made a confrontation of SVM and UMLFiT with bag-of-words features to classify the source of all calls on Brazilian Distrito Federal's official journal and conclude that SVM was still competitive with more modern methods.

Kowsari et al. (2019) identified a wide use of TF-IDF as feature for text classification and, as architecture, the use of deep, convolutional and deep belief neural networks, and Bi-LSTM. As good examples, Chen et al. (2018) used a DNN model with a 2D TF-IDF feature to classify Twitter comments concerning cyberbullying and hate speech. Finally, Chen et al. (2017) classified costumers reviews using a Bi-LSTM network followed by a 1D CNN with word embedding features.

3 Proposed Methodology

The proposed methodology is presented in Figure 1 and detailed on along with the next subsections. The workflow presented in Figure 1 is formed by two actors: Dataset and Classification, respectively.

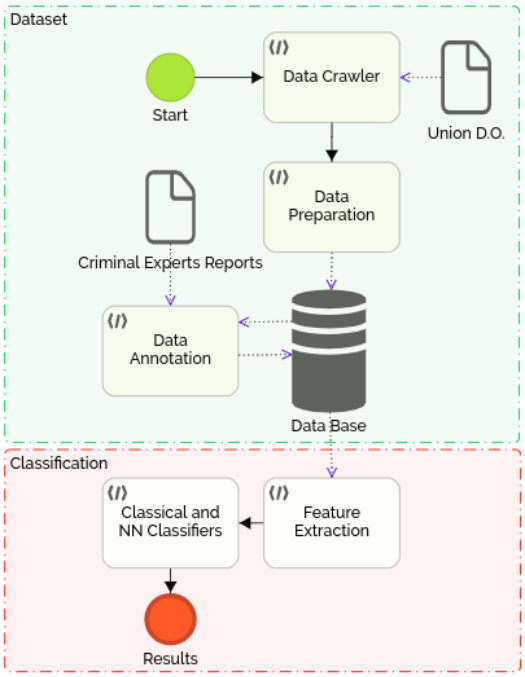


Figure 1: Workflow of the proposed methodology.

3.1 The Dataset Building Stage

The proposed dataset is a big set of text fragments extracted from DOU. All public procurement processed by the public administration uses this journal to make all its contracts. The Brazilian's laws no. 8666/1993 and no. 10520/2002 (Brasil, 1993, 2002) oblige all agencies to follow a strict set of rules for any agreement, and it is even more detailed for construction projects. By force of those Brazilian's laws, and by the constitutional publicity principle in this country, the main steps of the procurement process must be published as calls of application on the official media. Consequently, this data is accessible and reliable to serve as the main source of our proposed database.

Another indispensable characteristic of data most be its completeness (Weidema et al., 1996) and this is achieved knowing that, on Brazil, *DOU* Brazil (2020a) is a journal where all federal acts are publicized, it is divided into three sections and we can find on the third section: procurement calls, public tenders, contracts and it's addenda, public agreements, etc. There are several ways to obtain data of public procurements, as open tender systems, transparency portals. Still, in the field of Brazilian public works, they are spread by many sites and agencies and available in different formats, tables, documents, and detail levels amid the three levels of Brazilian State: federal, state, and cities.

Diário Oficial's publications, despite its relatively low level of detail, are very consistent and bring all vital information about the procurements as value, type, location, parts, object, etc.. They list without exception all public works of federal administration or with its direct financial support. This information raises the data reliability (Weidema et al., 1996) to be used by the criminal investigation and academic research.

3.1.1 Data Statistics

The database was obtained by a crawler algorithm developed by Ferreira (2018). It was applied to public data accessible at the site of Brazilian government enterprise for official publication as defined as *Imprensa Nacional* (Brazil, 2020a). A register sample of the dataset available for public download is shown in Figure 2.

Thus to form the database, the third section was downloaded from January 1998 to January 2018 into a PDF format and converted to a text format.

DIÁRIO OFICIAL DA UNIÃO

Figure 2: Sample register of the dataset.

Since February 2018, its publications are available on XML format and organized with a field for a public agency, type of document, etc. but without a specific field for procurement data, e.g. type of tender, deadlines, values, and scope. The dataset received information up to the last workday of February 2020 in a total of 15,132,968 entries. The Table 1 shows the dataset's primary data statistics.

Table 1: Dataset Statistics.

	Data
Entries	15,132,968
Max. length	1,040,513 characters or 156,703 words
Mean length	761.0 char. or 110.2 words

Publications were organized after the crawling stage on JSON files that include sequential identification, date, and raw text. After February 2018, tables are shown in text with HTML formatting. The length of the text field varies from a dozen to thousands of characters with titles of subsections and signatures names being the shortest and public tenders with names lists the longest. Although it is not a structured text, it maintains individual traces of order as it follows a formal and direct way of communication.

3.1.2 Data Annotation

The annotation of the public data was made with the use of a knowledge network of Brazilian forensics experts and do not represent an official assessment upon any person or public and private entity. The procurements, contracts, work, and/or agreements were annotated as having a fraud risk based on expert analysis. In their analysis its considered multiple indicators about date, place, type, parts (agency, enterprises), value, prices, execution, relation with other publications, and any other information linked to the process. So, it can not be concluded by a publication presence in this database about its legal or criminal status. Thereupon, we do not assess them as suspicion of fraud or not, but as a risk of fraud, and so publications were marked as having *risk = 1*.

All other publications included in the proposed dataset were marked, at first, as having *risk = 0*. Despite that, a procurement process is never said to be risk-free or fraud-free due to the nature of the criminal investigation (or an audit process). A total of 1,907 publications were marked as a risk of fraud, representing 0.012% of the dataset. The annotated data covers publications related to construction projects varying thousands to hundreds of million dollars in all Brazilian States.

As expected, the proposed dataset is very unbalanced due to the time demanded of an expert to fully analyze a public work procurement process. This results in a rate of 7,935 not annotated to 1 labeled as risk 1.

Instructions for downloading Deep-Vacuity dataset, go to URL: `http://www.cic.unb.br/~fbvidal/deepvacuity/dataset/index.html`.

3.2 The Classification Stage

The classification stage of this dataset tries to emulate a criminal expert assessment about the possibility of fraud in a given procurement. Experts, interviewed by the authors, said that value, agency, enterprises, location, date, type of construction and the correlation between that information usually lead to a good guess about the procurement risk of fraud. These gathered variables are the structured data models described in Section 2.

3.2.1 Training and Testing Subsets

To deal with the imbalanced dataset and successfully generalize the models, we created ten subsets with randomly chosen 1,907 entries of the not annotated publications to balance with annotated data. One way to classify the data is to consider all randomly chosen publications as having *risk = 0*. Although there is an error in that assumption, it can be assumed as low as the rarity of *risk 1* class. On top of that, the ten created subsets were divided into a training subset (with validation) and a test subset in ten-fold cross-validation archiving 100 training sets. Figure 3 illustrates how it was fulfilled.

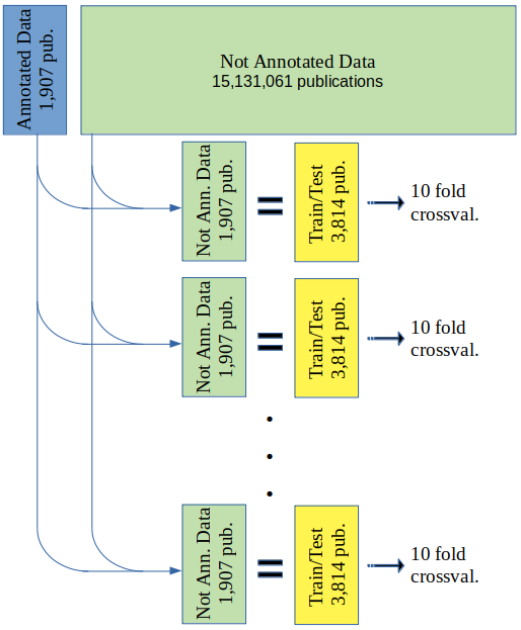


Figure 3: Training and Testing Subsets.

3.2.2 Comparison with Sparse Linear Classifiers

To create a baseline for comparisons, a wide range of classical linear supervised classifiers using sparse features was performed on the dataset, modeled using a Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction, that including the methods as follows:

- Stochastic Gradient Descent with Elastic Net with L1 Penalty (Zadrozny and Elkan, 2002; Zhang et al., 2002)
- Stochastic Gradient Descent with Elastic Net with L2 Penalty (Zadrozny and Elkan, 2002; Zhang et al., 2002)
- Linear SVC with L1-based feature selection (Fan et al., 2008)
- Naïve: Bernoulli, Complement and Multinomial (Manning et al., 2008; Rennie et al., 2003)
- Nearest Centroid (Tibshirani et al., 2002)
- Passive-Aggressive (Crammer et al., 2006)
- Perceptron (Freund and Schapire, 1999)
- Random Forest (Breiman, 2001)
- Ridge Classifier (Rifkin and Lippert, 2007)
- kNN with 10 neighbors (Altman, 1992)

The implementation of the Passive-Aggressive classifier method (Crammer et al., 2006) describes it as an online algorithm signifying that, for each prediction outcome for an instance of a sequential observation, the prediction mechanism is adjusted based on its correctness. A parameter of regularization controls this adjustment for the Passive-Aggressive method. Similarly, the Elastic Net penalty is one of the regularization parameters for the SGD method (Bottou, 2010). Figure 4 presents different F1 scores for the SDG using the parameters L1 and L2, being the Elastic Net, a compound of them, as defined by Zou and Hastie (2005). Those were the classifier methods with a higher F1 score (see Section 4). Benczúr et al. (2018) recommend an online learning strategy for scenarios with large data streams, because this sort of learning is based on each new event and its patterns.

3.2.3 Deep Neural Network Models

A current obstacle in science is to achieve good results using complex deep neural networks with few annotated data, but it is already a major advance in procurement fraud investigation, once deep neural networks use a supervised learning method.

For the feature extraction of the text, the TF-IDF approach was already used with linear classifiers and showed promising initial results. It was tested and implemented with deep neural networks, and initial tests showed better results. The vocabulary size was around 32 thousand words. The deep neural network models were build using the Tensorflow (Abadi et al., 2015). To evaluate the proposed dataset was chosen the Deep Neural Networks (DNN) and Bidirectional Neural Networks

models, all of them presented in Section 2 previously. Two DNN models were developed, the first model is defined as: $D(512, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(8192, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(512, 0.4, relu) \rightarrow D(256, 0.4, relu) \rightarrow D(1, 0, sigmoid)$, where $D(u, dr, a)$ indicates a dense layer with u nodes on output space with dr dropout rate and a as described as the activation function. The second is a bottleneck model defined as $D(8192, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow D(256, 0.4, relu) \rightarrow D(1024, 0.4, relu) \rightarrow D(2048, 0.4, relu) \rightarrow D(4096, 0.4, relu) \rightarrow D(8192, 0.4, relu) \rightarrow D(256, 0.4, relu) \rightarrow D(1, 0, sigmoid)$.

Many shapes and parameters of the **Bi-LSTM** model (Schuster and Paliwal, 1997) network were tested, but the one with better results was defined as: $BL(192, 0.25) \rightarrow BL(128, 0.25) \rightarrow BL(96, 0.25) \rightarrow TD(128, 0, relu) \rightarrow FL \rightarrow D(1024, 0, relu) \rightarrow D(2, 0, relu)$, where $BL(u, dr)$ is a Bi-LSTM layer with u and dr as defined above, TD is a time distributed dense layer similar to D, and FL is a layer that flattens the input. Outputs of the forward and backward RNNs were combined by concatenation in a length of 256 words.

4 Results

From every hundred tests ran for each model described above, it was computed precision, recall, and F1 Score with respective standard deviation. Those results, shown in Table 2, were plotted in Figure 4. The zoomed rectangle indicates most of the linear classifiers in a relative small F1 Score range from 91.4% to 93.4%. It can be observed that neural networks trended to produce a more significant standard deviation.

The best precision was achieved by the bottleneck network with 92.8%, the best recall was achieved by Naïve Complement with 97,2%, but the best value for F1 Score was by Passive-Aggressive with 93,4%.

Although the smaller F1-Score, both neural networks classifiers had higher precision, in other words, they classify less false positives entries. And knowing that the final model can raise a flag for further investigations and, on a finite resources police, that 1% difference equaling to an average monthly 569 entries, it is preferable to increase precision and lower recall than the opposite.

Table 2: Comparative Result of all trained models.

Method	**Mean**		
	Precision	**Recall**	**F1 Score**
D2V-SVM	90.1%	87.4%	88.7%
D2V-DNN	88.8%	85.1%	86.6%
GL.Bi-LSTM	91.0%	90.9%	90.9%
Bi-LSTM	91.4%	93.4%	**92.4%**
Bottleneck	92.8%	93.2%	**93.0%**
DNN	91.7%	93.2%	92.4%
Elastic-Net	90.2%	96.3%	93.2%
L1-penalty	89.6%	95.2%	92.3%
L2-penalty	90.1%	96.4%	93.1%
LinearSVC	88.6%	94.6%	91.4%
NaïveBern.	89.4%	96.6%	92.9%
NaïveCompl	86.2%	97.2%	91.3%
NaïveMulti	88.6%	96.3%	92.3%
NearestCen.	85.4%	90.4%	87.8%
Pass.-Agg.	90.7%	96.3%	**93.4%**
Perceptron	89.2%	95.4%	92.2%
Random-for.	88.6%	95.3%	91.8%
Ridge-Class.	89.7%	96.3%	92.9%
kNN	80.9%	95.5%	87.6%
DNNClass.	88.1%	90.6%	89.2%

The use of classical deep artificial neural networks classifiers proved that it is possible to use it on natural language processing applications to classify the procurement publications dataset and reach a reliable model to sort out risky procurements. Among the shelf feature extraction models, TF-IDF showed to be a better abstraction for the dataset and the better classical and neural networks classifiers obtained F1-Score results over 93%. Both bottleneck deep neural network and Bi-LSTM proved to be competitive with traditional classifiers and achieved better precision, which is more desirable (over recall) in a criminal fraud investigation.

Another contribution is towards in to use a basic set of LSTM models on a dataset without temporal correlation, as traditionally used by many other approaches. In the dataset used in our experiments, there is not any temporal correlation (or other features as dependency discourse parsing with word base, for example) among those collect data. Despite this feature, the used LSTM models achieved high performance, when compared with all baseline techniques. This issue opens new opportunities to possible developments using these models for solve situations when an application for text analy-

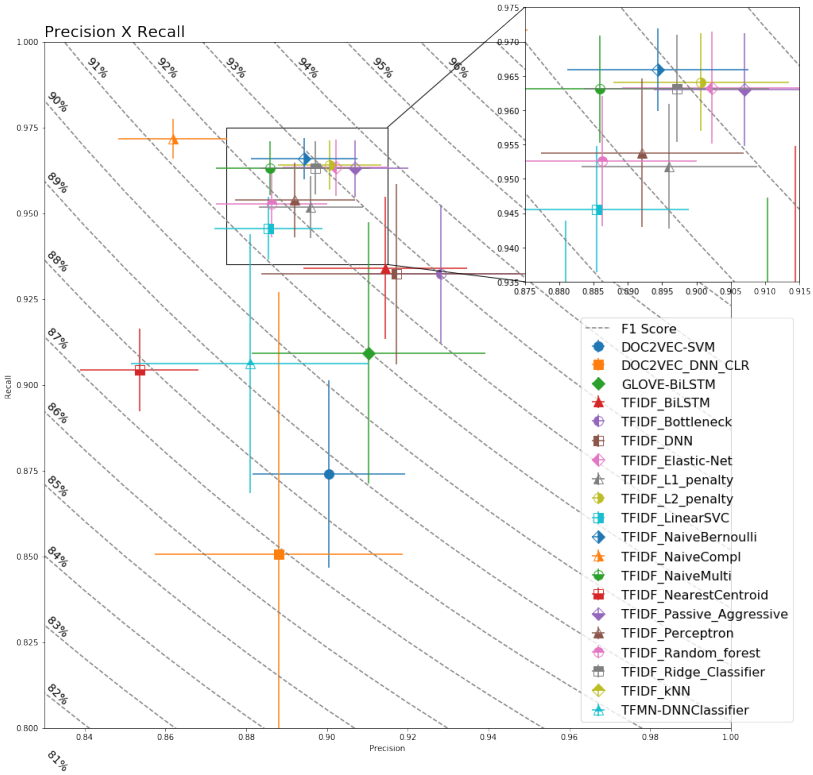


Figure 4: Precision, Recall and F1 Score of Models.

sis is achieved their limits (or bounded by boundary conditions), as a temporal correlation in the text fragment to be classified, for example.

5 Conclusions and Future Works

This work presented a preliminary evaluation of the DNN based on the LSTM model applied to fraudulent collusion detection from public works contracts in official texts. This proposed approach was compared with several state-of-the-art text classifiers (baseline classifiers), and the proposed method achieved competitive results for precision, recall, and F1 Score.

The proposed annotated dataset of the Brazilian public procurement calls allows researchers to explore a new form of fraud risk classification based on natural language processing and expert knowledge integration on its labeled data. The dataset covers already more than 22 years of publications, including mode than 15 million entries, and it will be available for academics researches.

On the other hand, despite the full range of customization of neural networks, it is possible to achieve even better results. It will be studied in the next works, new techniques to improve and customize feature extraction for the specific dataset and all LSTM models. For example, it is expected that Data Augmentation techniques should improve outcomes because of the small amount of annotated data available in the proposed dataset. Still, it is one of the many ways to achieve performance improvements in the classification process.

Acknowledgments

This work was supported by Brazilian Federal Police and University of Brasilia under the research project ”Pesquisa Aplicada de Inovações Tecnológicas no domínio da Perícia Criminal Federal”.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon

Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Jose María Alvarez, José Emilio Labra, Ángel Marín, and José Luis Marín. 2011. Semantic methods for reusing linking open data of the european public procurement notices. In *ESWC PhD Symposium*.

Jose María Alvarez-Rodríguez, Michalis Vafopoulos, and Juan Llorens. 2015. Enabling policy making processes by unifying and reconciling corporate names in public procurement data. the corfu technique. *Computer Standards & Interfaces*, 41:28 – 38.

Hubert Anysz, Andrzej Foremny, and Janusz Kulejewski. 2019. Comparison of ann classifier to the neuro-fuzzy system for collusion detection in the tender procedures of road construction sector. In *IOP Conference Series: Materials Science and Engineering*, volume 471, page 112064. IOP Publishing.

APCF. 2020. Perícia criminal. http://apcf.org.br/pericia-criminal/pericia-criminal. Visited: 2020-05-25.

Pedro Henrique Luz de Araujo, Teófilo Emidio de Campos, and Marcelo Magalhães Silva de Sousa. 2020. Inferring the source of official texts: can svm beat ulmfit? In *International Conference on Computational Processing of the Portuguese Language*, pages 76–86. Springer.

Remis Balaniuk, Pierre Bessiere, Emmanuel Mazer, and Paulo Cobbe. 2012. Risk based Government Audit Planning using Naïve Bayes Classifiers. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, Spain.

András A. Benczúr, Levente Kocsis, and Róbert Pálovics. 2018. Online machine learning in big data streams.

Leon Bottou. 2010. Stochastic gradient descent. https://leon.bottou.org/projects/sgd. Visited: 2020-05-24.

Brasil. 1993. Lei n. 8.666, de 21 de junho de 1993. http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm. Accessed: 2019-11-06.

Brasil. 2002. Lei n. 10.520, de 17 de julho de 2002. http://www.planalto.gov.br/ccivil_03/LEIS/2002/L10520.htm. Accessed: 2019-11-06.

Fabricio Ataides Braz, Nilton Correia da Silva, Teofilo Emidio de Campos, Felipe Borges S Chaves, Marcelo HS Ferreira, Pedro Henrique Inazawa, Victor HD Coelho, Bernardo Pablo Sukiennik, Ana Paula Goncalves Soares de Almeida, Flavio Barros Vidal, et al. 2018. Document classification using a bi-lstm to unclog brazil's supreme court. *arXiv preprint arXiv:1811.11569*.

Brazil. 2020a. Imprensa nacional. http://www.in.gov.br. Visited: 2020-04-23.

Brazil. 2020b. Painel de obras. http://transferenciasabertas.planejamento.gov.br/. Visited: 2020-05-19.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Ricardo Silva Carvalho and Rommel Novaes Carvalho. 2016. Bayesian models to assess risk of corruption of federal management units. In *BMA@ UAI*, pages 28–35.

Junyi Chen, Shankai Yan, and Ka-Chun Wong. 2018. Verbal aggression detection on twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, pages 1–10.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Hugo Honda Ferreira. 2018. Processamento de Linguagem Natural e Classificação de Textos em Sistemas Modulares. Monografia (Bacharel em Informática), Universidade de Brasília, Brasília, Brazil.

Open Knowledge Foundation. 2020. Procurement. http://index.okfn.org/dataset/procurement/. Visited: 2020-05-19.

Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

Sunita Goel and Ozlem Uzuner. 2016. Do sentiments matter in fraud detection? estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Kei Kawai and Jun Nakabayashi. 2014. Detecting large-scale collusion in procurement auctions. *Available at SSRN 2467175*.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Alan de Oliveira Lopes. 2011. Superfaturamento de obras públicas. *Livro Pronto, São Paulo, Brazil*.

Alan de Oliveira Lopes. 2015. O efeito pedagógico de operações da polícia federal: Um estudo de caso da operação" caixa de pandora". *Revista Brasileira de Ciências Policiais*, 6(1):67–85.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

OCDE. 2007. Bribery in public procurement methods, actors and counter-measures.

Fred Popowich. 2005. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1):59–66.

Kornelije Rabuzin and Nikola Modrušan. 2019. Prediction of public procurement corruption indices using machine learning methods.

Célia Ghedini Ralha and Carlos Vinícius Sarmento Silva. 2012. A multi-agent data mining system for cartel detection in brazilian government procurement. *Expert Systems with Applications*, 39(14):11642 – 11656.

Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623.

Ryan M Rifkin and Ross A Lippert. 2007. Notes on regularized least squares.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Prasad Seemakurthi, Shuhao Zhang, and Yibing Qi. 2015. Detection of fraudulent financial reports with machine learning techniques. In *2015 Systems and Information Engineering Design Symposium*, pages 358–361. IEEE.

Regis Signor, Peter ED Love, Alexanders TN Belarmino, and Oluwole Alfred Olatunji. 2020a. Detection of collusive tenders in infrastructure projects: Learning from operation car wash. *Journal of Construction Engineering and Management*, 146(1):05019015.

Regis Signor, Peter ED Love, Acir Oliveira Jr, Alan O Lopes, and Pedro S Oliveira Jr. 2020b. Public infrastructure procurement: Detecting collusion in capped first-priced auctions. *Journal of Infrastructure Systems*, 26(2):05020002.

Regis Signor, Peter ED Love, João José CB Vallim, Alexandre B Raupp, and Oluwole Olatunji. 2019. It is not collusion unless you get caught: the case of operation car wash and unearthing of a cartel. *Journal of Antitrust Enforcement*.

Ting Sun and Leonardo J. Sales. 2018. Predicting Public Procurement Irregularity: An Application of Neural Networks. *Journal of Emerging Technologies in Accounting*, 15(1):141–154.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

João José de Castro Baptista Vallim. 2020. Uso do Modelo de Raciocínio Baseado em Casos Para Monitoramento de Conluio em Licitações de Obras de Pavimentação Urbana. Master's thesis, Universidade Federal do Paraná, Curitiba.

John H Van Arkel, James J Wagner, Corrine L Schweyen, Saralyn M Mahone, Terrill J Curtis, Scott HAGINS, et al. 2013. Predictive modeling processes for healthcare fraud detection. US Patent App. 13/536,414.

Jonathan Watts. 2017. Operation car wash: Is this the biggest corruption scandal in history. *The Guardian*, 1(06):2017.

Weidema, Bo Pedersen, Wesnaes, and Marianne Suhr. 1996. Data quality management for life cycle inventories—an example of using data quality indicators. *Journal of cleaner production*, 4(3-4):167–174.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Tong Zhang, Fred Damerau, and David Johnson. 2002. Text chunking based on a generalization of winnow. *Journal of Machine Learning Research*, 2(Mar):615–637.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.