UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

IGOR SILVEIRA DE ANDRADE

USO DE PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZAGEM DE MÁQUINA PARA A EXTRAÇÃO DE INFORMAÇÃO EM EDITAIS DE LICITAÇÕES NÃO-ESTRUTURADOS

CAMPINA GRANDE - PB

2022

IGOR SILVEIRA DE ANDRADE

USO DE PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZAGEM DE MÁQUINA PARA A EXTRAÇÃO DE INFORMAÇÃO EM EDITAIS DE LICITAÇÕES NÃO-ESTRUTURADOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Professor Dr. Cláudio de Souza Baptista.

CAMPINA GRANDE - PB

2022

IGOR SILVEIRA DE ANDRADE

USO DE PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZAGEM DE MÁQUINA PARA A EXTRAÇÃO DE INFORMAÇÃO EM EDITAIS DE LICITAÇÕES NÃO-ESTRUTURADOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Cláudio de Souza Baptista
Orientador – UASC/CEEI/UFCG

Professor Dr. Maxwell Guimarães De Oliveira
Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de Setembro de 2022.

CAMPINA GRANDE - PB

ABSTRACT

Bidding is the means adopted by the public administration to ensure equality of conditions for all who want to contract products or services with the government. It is also the role of the government to carry out the analysis and audit of documents derived from this process, in order to guarantee legal principles such as isonomy, legality, impersonality, morality, publicity, and administrative probity. Most documents related to bidding processes use the Portable Document Format (PDF). Such an unstructured format makes automated textual analysis more complex. The present work aims to develop an induction model, based on supervised classification, that is able to identify specific information contained in a bidding document, and thus add a layer of automation to the document audit process. For this, natural language processing techniques were used, and different machine learning models were analyzed to select the best model to be used in the classification task. The database used was extracted from the Portal of the Government of the State of Acre. At the end of the implementation, the model obtained great results and was able to identify the information of interest present in the documents in a satisfactory way.

Uso de Processamento de Linguagem Natural e Aprendizagem de Máquina para a Extração de Informação em Editais de Licitações Não-Estruturados

Igor Silveira de Andrade
Laboratório de Sistemas de Informação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
igor.andrade@ccc.ufcg.edu.br

Cláudio de Souza Baptista
Laboratório de Sistemas de Informação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
baptista@computacao.ufcg.edu.br

RESUMO

A licitação é o meio adotado pela administração pública para assegurar igualdade de condições a todos que queiram realizar contratações de produtos ou serviços com o poder público. Também é papel do poder público, realizar a análise e auditoria dos documentos derivados desse processo, de forma a garantir princípios legais como isonomia, legalidade, impessoalidade, moralidade, publicidade, e probidade administrativa. Grande parte dos documentos relacionados aos processos licitatórios utilizam o formato Portable Document Format (PDF). Tal formato, não estruturado, torna a análise textual automatizada mais complexa. O presente trabalho, tem como objetivo o desenvolvimento de um modelo de indução, baseado em classificação supervisionada, que seja capaz de identificar informações específicas contidas em um edital de licitação, e assim adicionar uma camada de automação ao processo de auditoria do documento. Para isso, foram utilizadas técnicas de processamento de linguagem natural, e foram analisados diferentes modelos de aprendizagem de máquina, para a seleção do melhor modelo a ser utilizado na tarefa de classificação. A base de dados utilizada foi extraída do Portal do Governo do Estado do Acre. Ao final da implementação, o modelo obteve bons resultados e mostrou-se capaz de identificar as informações de interesse presentes nos documentos de maneira satisfatória.

Palavras-chave

licitação, auditoria, aprendizagem de máquina, NLP.

1 INTRODUÇÃO

O processo de licitação, ou processo licitatório, é um procedimento administrativo, isonômico, no qual a administração pública seleciona a proposta mais vantajosa, menos onerosa e que melhor atenda suas necessidades, para a contratação de uma obra, de um serviço, da compra de um produto, locação ou alienação [1]. Esses processos produzem uma enorme gama de documentos, planilhas e projetos, resultando em uma grande quantidade de informação gerada e não processada. O edital de licitação é um documento textual, não estruturado, no qual o órgão público determina as condições e exigências do processo licitatório [2]. Dentre os anexos de um edital de licitação, existe o termo de referência, ou projeto básico, que é um documento elaborado a partir dos estudos técnicos preliminares, que deve conter os elementos necessários e suficientes, com nível de precisão adequado, para caracterizar o objeto da licitação e outras informações importantes [3].

Um edital de licitação com conteúdo incompleto ou inconsistente não permite à administração pública realizar a seleção da proposta mais vantajosa, tendo, como consequência, o desperdício de recursos públicos. Por isso, faz parte do escopo dos órgãos públicos, como Tribunais de Contas dos Municípios, Estados e o da União, a análise e fiscalização das informações contidas em um edital de licitação, para garantir sua observância aos preceitos legais. Entretanto, a auditoria desses documentos não é uma tarefa trivial, esses longos documentos textuais comumente se apresentam no formato Portable Document Format (PDF), em forma de dados não-estruturados. Assim, normalmente, esse controle interno é feito através de uma verificação manual, onde o auditor de contas públicas, através da leitura do documento, deve averiguar a conformidade com a lei de tais editais. A partir desse problema, viu-se a necessidade e a oportunidade de adicionar um nível de automação a esse processo, através de um modelo indutor de aprendizagem de máquina, baseado em classificação supervisionada, capaz de classificar as sentenças do documento de acordo com as classes a ele apresentadas. Dessa maneira, a primeira etapa do processo de análise, a identificação das informações de interesse, poderá ser realizada de forma automática.

A extração de informações a partir de textos é uma das aplicações do Processamento de Linguagem Natural (PLN), um campo dentro da ciência da computação que busca converter a linguagem natural humana em uma representação formal, de modo que se torne mais facilmente manipulável por máquinas [4]. O grande desafio do PLN é transformar textos e falas, fontes de dados não estruturados, em conjuntos de dados capazes de serem lidos por uma máquina para o desenvolvimento de análises através de algoritmos de aprendizagem de máquina [5]. Atualmente, a análise de dados em linguagem natural utiliza diversos recursos para obtenção, interpretação, segmentação e categorização dos textos. Estes recursos têm contribuído para o progresso de sistemas automáticos e soluções inteligentes na análise de textos e são utilizados em diversos projetos recentes com alguns exemplos na língua portuguesa [6].

Neste trabalho, através de algoritmos de aprendizado de máquina, juntamente com técnicas de processamento de linguagem natural, buscou-se criar automaticamente classificadores de texto

por meio de um processo indutivo, utilizando-se de aprendizagem supervisionada, ou seja, de um corpus previamente anotado com as classes para realizar o treinamento e indução do modelo de aprendizagem. Desta forma, o modelo é treinado com uma referência de conhecimento e torna-se capaz de rotular diferentes tipos de textos, nunca por ele vistos. Esta abordagem traz uma precisão comparável à alcançada por especialistas humanos, economizando consideravelmente força de trabalho e diminuindo a necessidade de intervenção dos detentores do conhecimento de caso [7]. Para validação do modelo proposto foram utilizados editais de licitação reais, extraídos do Portal do Governo do Estado do Acre[1]. Os resultados alcançados foram excelentes, conforme as métricas de desempenho utilizadas e apresentadas mais adiante.

O restante deste artigo está estruturado da seguinte forma. Na seção 2 é apresentada a metodologia utilizada nesta pesquisa. Na seção 3, são discutidos os resultados obtidos. Por fim, na seção 4 são apresentadas as conclusões e apontamentos para futuros trabalhos.

2 METODOLOGIA

Para o desenvolvimento desta pesquisa, foi utilizada a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining), um modelo de processo que serve como base para um projeto de ciência de dados [8]. Esse modelo possui seis fases sequenciais que descrevem naturalmente o ciclo de vida de um projeto de mineração de dados, conforme indicado na Figura 1, são elas:

- *Entendimento do problema*: concentra-se na compreensão dos objetivos e requisitos do projeto, buscando referências e aprendendo mais sobre as regras de negócio;
- *Entendimento dos dados*: direciona o foco para identificar, coletar e analisar os conjuntos de dados que podem ajudar a atingir as metas do projeto;
- *Preparação dos dados*: prepara o conjunto de dados finais para modelagem. Nessa etapa acontece a seleção, limpeza, integração e formatação dos dados;
- *Modelagem*: testa e avalia diferentes modelos, com base em várias técnicas de modelagem;
- *Avaliação*: analisa mais amplamente qual dos modelos treinados atende melhor aos objetivos do negócio; e
- *Implantação*: implanta o melhor modelo escolhido a partir da análise das métricas de desempenho obtidas na etapa anterior de Avaliação. O modelo escolhido é implantado em produção, ou seja em um conjunto de dados reais, nunca vistos pelo classificador .[8].

2.1 Entendimento do problema

A análise de um edital de licitação, consiste na fiscalização das informações contidas no documento e seus anexos, de forma a garantir que estarão presentes todas as informações necessárias e obrigatórias ao processo e que não há inconsistências. Neste trabalho, o foco foi voltado para as informações presentes no corpo do edital e no termo de referência, um dos anexos do documento. Por se tratarem de longos documentos textuais, não estruturados, essa análise é feita de forma manual, aumentando a complexidade da tarefa. No presente estudo, foram utilizados dados do Portal do Governo do

[1]http://www.licitacao.acre.gov.br/editais/index.php

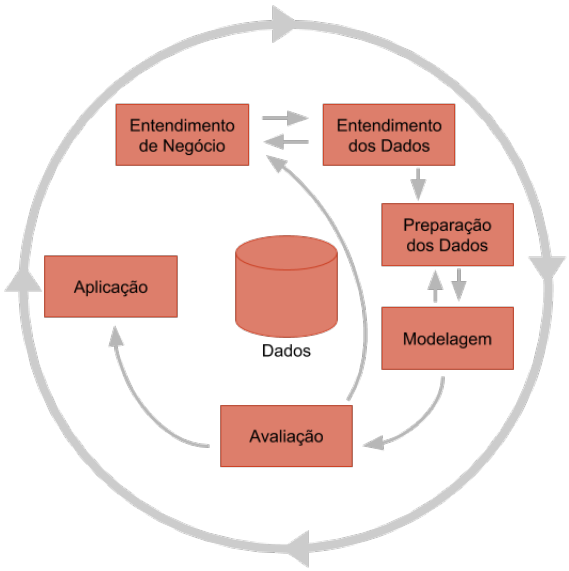

Figura 1: Etapas da metodologia CRISP-DM [9].

Acre, de onde foram baixados os editais de licitação, documentos textuais no formato PDF. Quando se trata desse formato, o primeiro desafio parte da extração de suas informações de maneira eficaz, para que se possa construir uma base de dados, maximizando a qualidade. Isso se dá, pois além da não estruturação dos dados em questão, os documentos muitas vezes apresentam imagens e tabelas que interferem e dificultam a extração das informações de maneira eficaz. Nessa etapa, foi definido que seria utilizada a linguagem Python, tendo em vista o seu vasto conjunto de bibliotecas para ciência de dados, aprendizagem de máquina e processamento de linguagem natural.

2.2 Entendimento dos dados

Os documentos extraídos, além do edital e do termo de referência, continham também outros anexos, como minuta de contrato, modelo de proposta de preços, ata de registro de preços, dentre outros, os quais não serão abordados neste trabalho. Os objetos de interesse desta pesquisa, o edital e termo de referência, contemplam diversas informações referentes ao processo licitatório, como objeto, condições para participação, habilitação, termos do contrato, entre outras. Após uma análise dos documentos e consulta aos Auditores de Contas do TCE-AC, foi definido um conjunto de informações, classes, para serem identificadas pelo classificador, foram elas:

- *Objeto*: declaração de modo conciso e completo, do que a Administração deseja contratar.
- *Local de entrega ou execução dos serviços*: local onde os produtos licitados devem ser entregues, e em caso de contratação de um serviço, onde ele será executado.
- *Qualificação técnica*: parte da etapa de habilitação, é o conjunto de requisitos profissionais que o licitante apresenta para executar o objeto da licitação.
- *Prazo de pagamento*: intervalo máximo de tempo para que seja realizado o pagamento referente ao objeto.

- *Prazo de entrega*: intervalo máximo de tempo para que seja realizada a entrega dos produtos licitados, e em caso de contratação de um serviço, para que o mesmo seja executado.
- *Cláusula de atraso de pagamento*: especificação das condições e multa aplicada em caso de eventuais atrasos de pagamentos.

Após a definição das informações a serem utilizadas e identificadas, foi feita uma inspeção no conjunto de dados para averiguar a qualidade dos mesmos. Foi observada a utilização de imagens e tabelas em meio aos documentos, o que poderia dificultar na extração dos textos presentes nesses itens e em seu entorno, por causarem uma quebra de formatação. Os problemas identificados foram tratados na etapa de preparação dos dados, descrita a seguir.

2.3 Preparação dos dados

2.3.1 Extração do texto.

A partir dos documentos PDF extraídos do Portal do Governo do Acre, fez-se necessária a extração do texto contido nos documentos para que posteriormente fossem rotulados os trechos contendo as informações de interesse (classes mencionadas anteriormente). Para a etapa de rotulagem, era necessário que as informações estivessem no formato de arquivo de texto (TXT), devido a compatibilidade da ferramenta a ser utilizada.

A primeira abordagem consistiu na tentativa de extração através da utilização de bibliotecas de manipulação de PDF, como PyMuPDF[2], PDFminer.six[3] e PyPDF2[4]. Os resultados obtidos, apesar de satisfatórios, ainda apresentaram algumas falhas devido às imagens e tabelas contidas nos documentos, onde em alguns casos, houve a perda ou corrupção das informações localizadas próximo a esses itens. O mesmo aconteceu com as informações contidas dentro das tabelas.

Na segunda abordagem, utilizou-se o Google Docs[5], uma ferramenta para edição de textos, planilhas e apresentações online disponibilizada gratuitamente pelo Google, para fazer a conversão do arquivo PDF para o formato DOCX. Devido aos recursos avançados utilizados pela ferramenta, o resultado dessa conversão foi muito fiel ao documento original. Partindo do documento no formato DOCX, que diferentemente do formato PDF, apresenta uma estruturação dos dados, e com o auxílio da biblioteca de manipulação de DOCX, python-docx[6], foi possível acessar separadamente as informações referentes a imagens, tabelas e parágrafos de maneira eficaz, e sem a perda ou corrupção das mesmas. Dessa maneira, foi possível acessar e extrair o texto contido nos documentos com ótima qualidade e assim criar os arquivos TXT.

Visto que os resultados da segunda abordagem foram mais satisfatórios, por não apresentarem as falhas obtidas na primeira abordagem, esse foi o método escolhido para realizar a extração do texto dos documentos no formato PDF, tendo como resultado final documentos no formato TXT, visando a etapa seguinte de rotulagem dos dados.

2.3.2 Rotulagem dos dados.

Foi realizada a rotulagem manual dos documentos com o auxílio do software Doccano[7], uma ferramenta de anotação de texto de código aberto que fornece recursos de anotação para classificação de texto, rotulagem de sequência e tarefas de sequência a sequência. Foram definidos *labels* de acordo com as classes mencionadas na subseção 2.2, e caso o conteúdo da sentença contenha a informação de interesse, a mesma é classificada com o respectivo *label*, como pode ser observado na Figura 2 , sendo assim, uma classe positiva.

O *corpus* resultante das classes positivas tem uma estrutura semelhante a da Tabela 1.

Sentença	**Classificação**
Contratação de Empresa para Fornecimento de Leito Filtrante para ETAS Compactas e Abertas destinados a atender as necessidades do Departamento Estadual de Água e Saneamento – DEPASA.	Objeto
O objeto da licitação deverá ser entregue na usina do DERACRE, situado na Rua Topógrafo Domingos nº 511 – Distrito industrial, CEP 69.901-180, Rio Branco – Acre.	Local de entrega ou execução dos serviços
Atestado de capacidade técnica, expedido por pessoas jurídicas de direito público ou privado, que comprovem ter o licitante fornecido satisfatoriamente os serviços pertinentes e compatíveis com o objeto desta licitação.	Qualificação técnica
O Prazo de Entrega dos medicamentos será de até 25 (vinte cinco) dias consecutivos contados a partir da data de emissão de Ordem de Entrega emitida pela Central Demandante.	Prazo de entrega
O pagamento será efetuado até o 30º (trigésimo) dia subsequente à entrega dos materiais, mediante apresentação da nota fiscal devidamente atestada por servidor responsável.	Prazo de pagamento
Quando da ocorrência de eventuais atrasos de pagamento [...] mediante aplicação da seguinte fórmula: EM = VP x N x I, onde: EM = Encargos moratórios; VP = Valor da parcela em atraso; N = Número de dias entre a data prevista para o pagamento e a do efetivo pagamento; I = Índice de atualização financeira; TX = Percentual da taxa de juros de mora anual;	Cláusula de atraso de pagamento

Tabela 1: Exemplo de classificação das sentenças do texto.

Para as sentenças que não foram classificadas com nenhuma das seis classes de interesse mencionadas na subseção 2.2, foi criada a classe *outros*, sendo essa a classe negativa. Ao final do processo de rotulagem foram obtidas 18.488 sentenças, onde as classes positivas

[2]https://pymupdf.readthedocs.io/
[3]https://pdfminersix.readthedocs.io/
[4]https://pypdf2.readthedocs.io/
[5]https://www.google.com/intl/pt-BR/docs/about/
[6]https://python-docx.readthedocs.io/en/latest/
[7]https://github.com/doccano/doccano

ocuparam 11% dos registros e a classe negativa 89%. Esse desbalanceamento entre a quantidade de registros entre as classes positivas e a negativa era esperado, visto que em um documento real, a quantidade de sentenças contendo alguma das informações de interesse é muito menor quando comparada com o total de sentenças do documento.

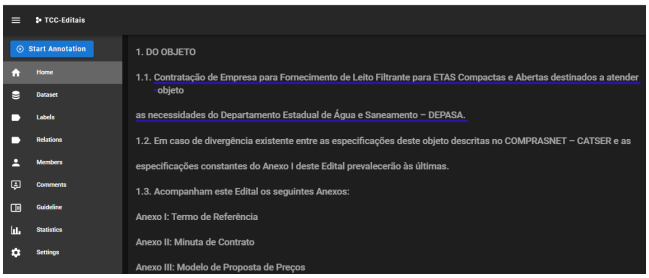

Figura 2: Exemplo de uso do Doccano.

2.3.3 Pré-processamento dos dados.
Com os dados coletados, foram então removidas as sentenças repetidas, fazendo com que a quantidade de dados caísse para 14.137 sentenças diferentes, onde as classes positivas ocuparam 12% dos registros e a classe negativa 88%. Os dados foram submetidos à técnica de capitalização, isto é, a normalização do texto em uma só tipografia, caixa alta ou baixa. Outra técnica utilizada foi a lematização, que tem como objetivo reduzir uma palavra à sua forma base e agrupar diferentes formas da mesma palavra [10]. Por fim, também foram aplicadas as técnicas de pré-processamento: remoção de stopwords e tokenização, fazendo uso da biblioteca nltk[8] e do algoritmo TfidfVectorizer da biblioteca sklearn[9], respectivamente. É importante ressaltar que para o modelo baseado em transformer, BERTimbau, não foi realizado nenhum pré-processamento, conforme indicado pela literatura.

2.4 Modelagem

Os modelos de classificação escolhidos para serem analisados podem ser divididos em dois grupos: modelos de classificação tradicionais e modelo utilizando aprendizagem profunda. Os algoritmos tradicionais utilizados foram: *Decision Tree*, *Support Vector Machine (SVM)* e *XGBoost*. O modelo de aprendizagem profunda selecionado foi o transformer BERTimbau[10], um modelo BERT pré-treinado para o português brasileiro que alcança desempenho de última geração [11].

O conjunto de dados foi dividido em treino, teste e validação, onde, de cada classe, selecionados de forma aleatória, 20% dos dados foram destinados ao conjunto de teste, 10% ao conjunto de validação e 70% destinados para o conjunto de treinamento.

Ao fim da divisão entre os conjuntos, o conjunto de treinamento apresentou um grande desbalanceamento entre o número de sentenças de cada classe, como pode ser observado na Figura 3. Por isso, para balancear esse conjunto, utilizou-se de uma técnica de aumento de dados conhecida como *back translation*. Essa técnica consiste em três etapas principais: *Tradução temporária*, em que cada um dos dados de treinamento original foi traduzido para um idioma diferente, no nosso caso do português para o francês; *Retrotradução*: foi traduzido de volta cada um desses dados traduzidos para o idioma original, ou seja, uma tradução do francês para o português; e por fim, *Remoção de duplicatas*, onde ao final do processo, foi removido todas as sentenças duplicadas [12]. Através dessa técnica, foi possível criar dados artificiais, com qualidade semelhante aos dados anotados, e assim, balancear o conjunto de treinamento. Após o balanceamento, o conjunto de treinamento ficou composto de 1.200 sentenças das classes positivas, igualmente dividido entre as classes, e 1.200 sentenças da classe negativa.

Nos modelos de classificação, para reduzir as chances dos modelos sofrerem de *overfitting*, foi utilizado o método de validação cruzada *k-folds*, com o parâmetro k = 10, através do algoritmo *StratifiedKFold* disponibilizado pelo sklearn. Para automatizar o processo de ajuste dos hiperparâmetros dos modelos criados durante o processo de treinamento, foi utilizado o algoritmo *GridSearchCV* do sklearn. Os valores dos parâmetros utilizados nos modelos estão detalhados na Tabela 2.

Para indução do modelo BERTimbau, que se trata de um modelo de linguagem pré-treinado para língua portuguesa, utilizando-se de uma arquitetura de rede neural profunda conhecida como *Transformers*, foram utilizados tensores *pytorch* [11] durante 4 épocas na fase de *fine tuning*, valor este dentro dos limites recomendado pela literatura, utilizando o mesmo conjunto de treinamento dos modelos do sklearn.

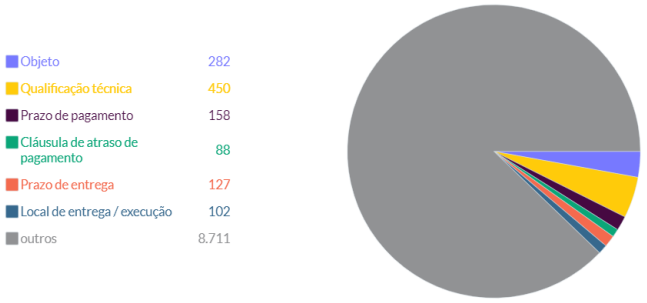

Figura 3: Distribuição inicial do conjunto de treinamento.

Modelo	Parâmetro	Valor
XGBoost	max_depth	4
	colsample_bytree	0.6
	subsample	0.8
	min_child_weight	1
SVM	C	100
	gamma	0.01
	kernel	rbf
Decision Tree	criterion	entropy
	max_depth	40

Tabela 2: Parâmetros utilizados nos modelos tradicionais.

[8] https://www.nltk.org
[9] https://scikit-learn.org/stable/user_guide.html
[10] https://github.com/neuralmind-ai/portuguese-bert
[11] https://pytorch.org/docs/stable/index.html#pytorch-documentation

3 AVALIAÇÃO DE RESULTADOS

As métricas utilizadas para avaliar os modelos, medindo sua eficácia diante do problema, foram: acurácia, precisão, recall e F1-score. Essas métricas são calculadas a partir dos dados obtidos da matriz de confusão: verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN), como consta na Figura 4. Após encontrar os melhores hiperparâmetros dos modelos selecionados, foram obtidas as seguintes métricas, listadas na Tabela 3 baseadas no conjunto de testes.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precisão} = \frac{TP}{TP + FP} \qquad \text{F1-score} = 2 * \frac{\text{Precisão} * \text{Recall}}{\text{Precisão} + \text{Recall}}$$

Figura 4: Equações das métricas utilizadas.

Modelo	Métrica	Valor
Decision Tree	acurácia	93,23%
	precição	72,39%
	recall	98,15%
	f1-score	79,04%
XGBoost	acurácia	95,98%
	precição	85,80%
	recall	86,91%
	f1-score	86,18%
SVM	acurácia	96,75%
	precição	86,50%
	recall	93,84%
	f1-score	89,88%
BERTimbau	acurácia	97,00%
	precição	97,33%
	recall	97,00%
	f1-score	97,10%

Tabela 3: Resultados das métricas dos modelos.

Com base nas métricas obtidas, pode-se observar que o modelo pré-treinado BERTimbau obteve os resultados mais satisfatórios, superando os outros modelos em todas as métricas calculadas. A matriz de confusão da Figura 5 mostra sua performance no conjunto de dados de teste, conjunto esse em que foi mantida a proporção original dos dados, e a quantidade de sentenças de cada classe reflete a dos documentos reais, em que as sentenças da classe negativa são a grande maioria. Na matriz de confusão, as classificações apresentadas como *0*, *1*, *2*, *3*, *4*, *5* e *6*, são equivalentes às classes *Objeto*, *Qualificação Técnica*, *Prazo de pagamento*, *Cláusula de atraso de pagamento*, *Prazo de entrega*, *Local de entrega ou execução dos serviços* e *outros*, respectivamente. Dessa forma, pode-se concluir que o BERTimbau foi o melhor modelo na tarefa de classificar as sentenças de um documento de edital de licitação, entre as classes apresentadas ao modelo.

Para verificar a ausência de *overfitting*, foi gerado o gráfico da Figura 6, representando o comportamento do erro durante o treinamento com os dados de treino e validação. Neste gráfico, podemos observar que o erro com dados de validação vai caindo à medida que o de treinamento também cai. Logo, podemos concluir que o modelo está generalizando bem.

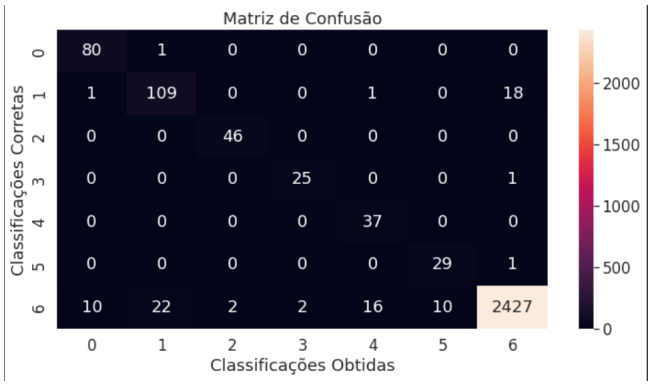

Figura 5: Matriz de confusão.

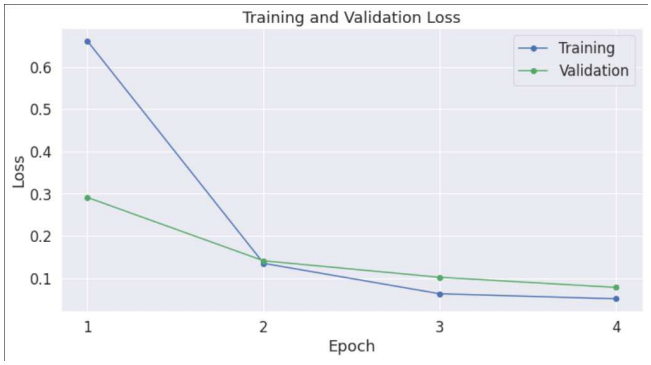

Figura 6: Gráfico do erro.

4 CONCLUSÃO

O presente trabalho buscou criar um modelo inteligente capaz de classificar, dentre um conjunto de classes predefinidas, as sentenças de um documento de edital de licitação e em seu anexo, o termo de referência, com o intuito de auxiliar na auditoria desses documentos. Para isso, foi feita uma análise entre diversos modelos de classificação, onde o BERTimbau, modelo pré-treinado baseado em transformers, obteve os melhores resultados, sendo eleito o mais apto a realizar a tarefa objetivada. Tendo em vista trabalhos futuros, pretende-se testar outros modelos classificadores de arquitetura transformers para se comparar com os resultados obtidos. Modelos mais atuais, como o *Open Pre-trained Transformer (OPT)* [13] e o *Generative Pre-Training Transformer 3 (GPT-3)* [14], são exemplos de modelos de aprendizagem profunda com capacidade para até 175 bilhões de parâmetros de aprendizagem de máquina, e atualmente se apresentam como estado da arte.

AGRADECIMENTOS

Agradeço à minha família, por todo amor e suporte, em especial à minha mãe, Suely de Fátima, e minha tia, Susete Maria, por todos os sacrifícios feitos por mim para que eu pudesse seguir focado no caminho acadêmico. À minha namorada, Luiza Maria, por todo apoio e por estar sempre ao meu lado de forma incondicional. Ao Professor Cláudio Baptista, meu orientador, pelas oportunidades e pela confiança durante todo esse tempo de convivência. Aos meus amigos, por toda ajuda, conselhos e momentos compartilhados, em especial à Gileade Kelvin, Jéssica Gabriele, Rich Elton, Matheus Santana, Samuel Vasconcelos, Vinicius Barbosa, José Davi, Yuri Souza e Levi Gomes. Por fim, agradeço a todos os professores e funcionários da Universidade Federal de Campina Grande, que de forma direta ou indireta, contribuíram para meu crescimento acadêmico e profissional.

REFERÊNCIAS

[1] LICITAçõES e contratações - Portal da transparência. 2022. https://www.portaltransparencia.gov.br/entenda-a-gestao-publica/licitacoes-e-contratacoes. Acesso em: 22 de mar. de 2022. 1

[2] EDITAL de Licitação. 2022. https://www.licitacao.net/edital_de_licitacao.asp. Acesso em: 22 de mar. de 2022. 1

[3] TERMO de referência ou projeto básico. 2022. http://www.tcu.gov.br/arquivosrca/001.003.011.htm. Acesso em: 18 de jul. de 2022. 1

[4] MANNING, C. D. et al. The stanford corenlp natural language processing toolkit. In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 55–60. 1

[5] JURAFSKY, D.; MARTIN, J. *Speech and Language Processing*. Pearson Education, 2014. ISBN 9780133252934. Disponível em: <https://books.google.com.br/books?id=Cq2gBwAAQBAJ>. 1

[6] FERNANDES, R. V. D. C.; COSTA, H. A.; CARVALHO, A. G. P. D. Tecnologia jurídica e direito digital. In: *I Congresso Internacional de Direito e Tecnologia*. [S.l.]: Editora Forum, 2018. 1

[7] SEBASTIANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 34, n. 1, p. 1–47, mar 2002. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/505282.505283>. 2

[8] WHAT is CRISP DM? *Data Science Process Alliance*, c2022. Disponível em: <https://www.datascience-pm.com/crisp-dm-2/>. Acesso em: 01 de mar. de 2022. 2

[9] MOURA, K. *Ciclo de vida dos dados 2*. Medium, 2019. Disponível em: <https://medium.com/@kvmoura/crisp-dm-79580b0d3ac4>. Acesso em: 01 de set. de 2022. 2

[10] YSE, D. L. *Your guide to natural language processing (NLP)*. Towards Data Science, 2019. Disponível em: <https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1>. 4

[11] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. 4

[12] FADAEE, M.; BISAZZA, A.; MONZ, C. Data augmentation for low-resource neural machine translation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 567–573. Disponível em: <https://aclanthology.org/P17-2090>. 4

[13] ZHANG, S. et al. *OPT: Open Pre-trained Transformer Language Models*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2205.01068>. 5

[14] BROWN, T. B. et al. *Language Models are Few-Shot Learners*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>. 5