Descoberta de tópicos e classificação de textos de licitações promovidas pelos órgãos públicos do estado da Bahia

Gabriel Andrade de Sant'Anna
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
gabriel.santanna@mpba.mp.br

Ludmilla Palmeira Andrade
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
ludmilla.andrade@mpba.mp.br

Rubia Teles de Souza
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
rubia.souza@mpba.mp.br

Braian Varjão Gama Bispo
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
braian.bispo@fieb.org.br
braiangama@gmail.com

***Resumo*—As compras e contratações dos Órgãos Públicos são realizadas através de processos licitatórios, que tem como objetivo cumprir suas funções sociais, respeitando princípios da legalidade, da impessoalidade, da moralidade, da publicidade, do interesse público, da probidade administrativa, dentre outros [1]. O controle desses processos licitatórios auxilia a gestão adequada dos recursos públicos. Com o objetivo de melhorar a análise sobre as licitações, o Ministério Público do Estado da Bahia desenvolveu o BI – Licitômetro. Essa ferramenta possui entre as suas funcionalidades de análise a capacidade de agrupar as licitações por categorias através do uso das palavras chaves. Entretanto, o método aplicado para a caracterização das licitações não tem demonstrado eficiência, gerando classificações incorretas e dificultando a análise das licitações. Esse trabalho propõe a utilização do BERTopic para realizar a modelagem de tópicos das licitações efetuadas no Estado da Bahia junto com uma abordagem supervisionada de classificação e a utilização de um método não-supervisionado para classificação das licitações referentes aos serviços de assessoria/consultoria jurídica. O objetivo é comparar os resultados dos métodos de classificação e selecionar o melhor para implantar no BI-Licitometro. O método não supervisionado de classificação utilizado foi o Zero-Shot Classification e o método supervisionado foi o Random Forest Classification.**

***Palavras-chave*—classificação de texto; modelagem de tópicos; BERTopic, Zero-Shot Classification; Licitações.**

I Introdução

Em Órgão Público, as compras, locações, concessões e permissão de uso de bens públicos, prestações de serviços (incluindo os técnico-profissionais especializados), contratações de obras de arquitetura e engenharia e de tecnologia da informação e comunicação precisam passar por procedimentos administrativos formais chamados de Licitações [2] [1]. Esses procedimentos devem obedecer às normas gerais da Lei 14.133, de abril de 2021, observando alguns princípios, como: legalidade, impessoalidade, moralidade, publicidade, interesse público, probidade administrativa, dentre outros [1]. Esta Lei visa assegurar que a Administração Pública selecione a proposta mais vantajosa para suas contratações, além de proporcionar aos licitantes um tratamento isonômico e com justa competição [2]. Desta forma, todo o processo que envolve as contratações pelo serviço público está sujeito às fiscalizações da sociedade civil, dos Tribunais de Contas e dos Ministérios Públicos estaduais e federais. A Lei 12.846, de 1º de agosto de 2013, sobre a responsabilização administrativa e civil de pessoas jurídicas pela prática de atos contra a administração pública, diz que frustrar ou fraudar procedimentos licitatórios constitui ato lesivo à administração pública e que o Ministério Público poderá ajuizar ação visando aplicações de sanções às pessoas jurídicas infratoras [3]. Segundo o art. 129 da Constituição Federal, é função do Ministério Público

promover inquérito civil para proteção do patrimônio público. As fiscalizações objetivam identificar irregularidades, o mau uso do dinheiro público, além de embasar investigações de improbidade administrativas e atos corruptos que podem causar grandes prejuízos sociais. Com o objetivo de apoiar a análise das compras e aquisições do estado da Bahia, o Ministério Público Estadual desenvolveu o BI Licitômetro. Esta ferramenta possibilita o desenvolvimento de várias análises sobre os processos de compras firmados pelos órgãos públicos do estado da Bahia. Uma das funcionalidades do Licitômetro é agrupar as licitações através da descrição dos objetos dos processos de compra. Atualmente esse agrupamento é feito por palavras-chave, o que tem gerado incoerências na determinação das classes dos objetos licitados, pois a mesma palavra-chave está presente em muitas descrições de objetos distintos, dificultando a classificação desses itens, já que o objeto licitado pode não corresponder ao que está sendo classificado pelo filtro. Com o intuito de aperfeiçoar esse filtro, diversas técnicas tradicionais de tratamento de dados foram utilizadas, porém nenhuma foi satisfatória para a solução desse problema. Nessas experiências, alguns obstáculos foram notados, como: o grande volume de licitações sem classificação de objeto e a flexibilidade existente na descrição dos textos, que dificulta uma classificação mais aderente ao que foi contratado pelos órgãos públicos. Vale salientar que a base de dados de licitações possui mais de um milhão de registros. A cada ano são realizadas em torno de 200.000 licitações no estado da Bahia. Os temas dessas contratações são os mais variados possíveis, podendo ser a compra de copos plásticos até a contratação de empresas para a construção de unidades de saúde. No âmbito das licitações públicas, as descrições dos objetos que determinam as contratações realizadas pelo estado podem ser cruciais para a análise dos gastos públicos, a identificação de mau uso das verbas públicas, apoio às investigações, dentre outras tarefas referentes às fiscalizações do estado. Diante dessa realidade e com a finalidade de agrupar as licitações de acordo com as descrições dos objetos licitados, este trabalho tem como objetivo realizar um estudo comparando modelo supervisionado com modelo não supervisionado para classificar as licitações referentes aos serviços de assessoria/consultoria jurídica. Para o treinamento supervisionado de Machine Learning, o método escolhido foi o Random Forest Classification. Aliado a ele, foi aplicada a técnica de modelagem de extração de tópicos para a criação do dataset, utilizando o resultado dos tópicos extraídos para determinar os rótulos dos objetos. Utilizou-se o método Zero-Shot Classification para o modelo não supervisionado, que apresentou o resultado mais eficaz no estudo com grande potencial de ser utilizado no BI. Nas próximas seções serão abordados os seguintes tópicos: referencial teórico – uma síntese das tecnologias utilizadas no estudo; metodologia do trabalho – o passo a passo do experimento; resultados obtidos e trabalhos futuros.

II Referencial Teórico

A caracterização de textos através do acesso ao seu conteúdo é um dos grandes problemas nas áreas de processamento de linguagem natural (PNL) e de aprendizagem de máquina. Frequentemente, esta caracterização dará base para as tarefas de recuperação, classificação e previsão de textos [4]. A descoberta dos temas das licitações é uma aplicação direta da modelagem de tópicos. Essa tarefa tem como objetivo encontrar grupos de palavras (tópicos) em um conjunto de textos com o objetivo de identificar os principais assuntos presentes nestes documentos utilizando uma abordagem não supervisionada de modelo de machine learning [5]. A extração de tópicos será a atividade que viabilizará o treinamento de modelos para classificação das licitações futuras. A seguir serão apresentados os conceitos e tecnologias que foram utilizados na construção deste trabalho.

A. Inteligência Artificial, Aprendizado de Máquina e Aprendizado Profundo

A inteligência artificial é o campo da ciência que busca automatizar através de máquinas a atuação de seres humanos na realização de atividades com características que dependem da inteligência humana, ou seja, é a ciência que estuda e constrói máquinas inteligentes. O Aprendizado de Máquina (Machine Learning) é uma subárea da Inteligência Artificial que tem foco no aprendizado de sistemas inteligentes; já o Aprendizado Profundo (Deep Learning) é uma subárea do Aprendizado de Máquina

que aplica soluções através da utilização de redes neurais artificiais [6] [7].

B. Tipos de Aprendizado de Máquina

O Aprendizado de Máquina é o desenvolvimento de programas que obtém melhoras no seu desempenho a partir de exemplos [8]. Esses exemplos são obtidos através da obtenção de grandes quantidades de dados para a construção do conhecimento do computador, tornando possível o aprendizado por dados [9] , ao invés do uso exclusivo de instruções pré-programadas [10]. O principal objetivo do aprendizado de máquina é gerar um modelo de predição, classificação ou detecção [11]. O Aprendizado de Máquina depende da intervenção humana para desenvolver os algoritmos, pois existe a necessidade de avaliar o contexto para aplicação das técnicas que viabilizem a construção de um modelo que possa resolver o problema de maneira satisfatória. As técnicas são normalmente baseadas no princípio indutivo. Nessa abordagem uma conclusão genérica é definida a partir de um conjunto específico de treinos. Os tipos de aprendizado de máquina podem ser classificados em três subgrupos de algoritmos indutivos: supervisionados - nesse tipo de algoritmo, o conjunto de dados é previamente rotulado e conhecido; semi-supervisionado - nesta abordagem, uma porção dos dados estão rotulados e é possível usar esta informação para ajudar no processo de agrupamento e identificação dos registros não rotulados; não-supervisionados - o algoritmo tenta descobrir os padrões e estruturas para organizar e agrupar os dados [12].

C. Transformers e Modelos de Linguagem

Os transformers são algoritmos baseados em redes neurais com capacidade de executar tarefas de processamento de linguagem natural (PLN). Os modelos de linguagem baseados neles são utilizados entre outras soluções para a tradução automática, geração de legendas e resumo de textos. A arquitetura desses modelos é denominada transformer [5]. Os modelos de linguagem baseados em transformers são treinados em duas etapas principais: pré-treinamento e ajuste fino (fine-tuning). No pré-treinamento, o modelo recebe como entrada textos para uma tarefa de previsão de termos ausentes. Aprendendo a prever termos ocultos (mascarados) em uma sequência de palavras. Após a etapa de pré-treinamento, é iniciada a fase do fine-tuning, momento em que o modelo é ajustado para uma tarefa específica, como geração ou tradução de textos a partir de um conjunto de dados rotulados, fazendo com que o modelo seja capaz de produzir resultados mais relevantes e precisos [5]. O modelo de representação de linguagem pré-treinado BERT (Bidirectional Encoder Representations from Transformers) foi desenvolvido pelo Google. Ele é baseado na arquitetura Transformer [13], tem como característica a capacidade de tratar sequências de textos de comprimento variável, paralelismo e uma melhor eficiência de tempo em relação a técnicas anteriores [14]. O BERT possibilitou a criação de variantes para diversos idiomas, como os modelos: RoBERTA e BERTimbau para o idioma português, e o XLM-RoBERTA que é uma versão multilíngue treinada com dados de 100 idiomas [5]. Os modelos BERT proporcionaram a criação de outras abordagens para o processamento de linguagens naturais, entre elas o BERTopic.

D. Modelagem de Tópicos

O BERTopic é um algoritmo desenvolvido para a extração de tópicos a partir de um grande conjunto de documentos [15]. Essa técnica utiliza a representação semântica fornecida pelo BERT, extraindo os embeddings dos textos do modelo de linguagem e realizando o agrupamento desses embeddings em clusters de tópicos representativos. O BERTopic tem se destacado em relação aos modelos mais tradicionais de extração de tópicos, gerando representações mais ricas, contextuais e coerentes [5]. Esse algoritmo considera as características semânticas das representações vetoriais, permitindo a codificação do significado do texto e aproximando os documentos semelhantes no espaço vetorial [16].

E. Classificação de Textos Através do Aprendizado de Máquina

A classificação de textos é uma das aplicações mais difundidas na área de processamento de linguagem natural (PLN). Através desses tipos de soluções é possível identificar as categorias de artigos jornalísticos (política, culinária, saúde etc.) ou classificar um comen-

tário de um blog como positivo ou negativo [17]. Essa técnica pode ser aplicada de forma supervisionada ou não-supervisionada.

F. Zero-Shot Classification

O Zero-shot classification é uma técnica de aprendizado de máquina capaz de classificar dados em categorias que não foram vistas durante o treinamento. Diferente dos métodos tradicionais de aprendizado supervisionado, que requerem grandes quantidades de dados rotulados para cada categoria, o zero-shot classification permite que o modelo generalize para novas classes, utilizando apenas descrições semânticas ou relacionamentos das classes.[18] Este modelo baseia-se na sua compreensão geral do mundo, linguagem, conceitos ou padrões para fazer previsões ou decisões sobre estas novas categorias, por exemplo: se uma pessoa conhece um cavalo, mas nunca viu uma zebra, poderia reconhecer uma sabendo que a zebra parece um cavalo com listras pretas e brancas. O Zero-Shot Classification consegue de maneira similar ao exemplo descrito anteriormente reconhecer uma relação semântica entre classes já conhecidas e ainda não vistas [19].

G. Métricas para Avaliação dos Modelos

As métricas dos algoritmos de machine learning são números que possibilitam a avaliação dos modelos de maneira quantitativa [20]. Este trabalho utilizou a Coerência como métrica para avaliação de tópicos, por ser a mais similar ao julgamento humano [21] [22]. Na análise com as áreas de negócios foi avaliado que a acurácia e a precisão são as métricas mais aderentes para avaliação dos modelos de classificação utilizados na solução do problema. Pois, a acurácia informa o quanto as licitações foram classificadas corretamente, independentemente da classe. Já a Precisão enfatiza o quanto o modelo acerta quando classifica um tipo como positivo. Além disso, os dados utilizados no treino apresentaram balanceamento, reduzindo os vieses na avaliação dos modelos. A Acurácia é o resultado da divisão das previsões corretas pelo total de previsões realizadas.

$$\text{Acurácia} = \frac{\text{Número de Previsões Corretas}}{\text{Total de Previsões}}$$

A Precisão é a divisão da quantidade de previsões positivas e corretamente classificadas pelo total de previsões positivas do modelo.

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos (VP)}}{\text{Verdadeiros Positivos (VP)} + \text{Falsos Positivos (FP)}}$$

Uma ferramenta que foi utilizada para melhorar a observação das métricas do modelo de classificação foi a Matriz de Confusão. A matriz de confusão é uma tabela que organiza os acertos e erros do modelo treinado colocando a classificação real dos registros nas linhas horizontais e os valores preditos pelo modelo nas linhas verticais. Através desta organização é possível extrair os verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos(TN) e falsos negativos (FP). A seguir é apresentada uma representação de uma matriz de confusão.

Tabela I: Matriz de Confusão

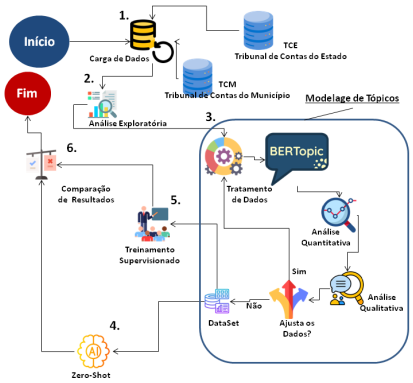
		Condição Preditiva	
	População Total = P + N	Positivo (PP)	Negativo (PN)
Condição Atual	Positivo (P)	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Negativo (N)	Falso Positivo (FP)	Falso Negativo (FN)

Fonte: Dados da pesquisa.

III METODOLOGIA

O estudo foi dividido em 6 etapas: seleção e carga de dados (1), análise exploratória dos dados (2), modelagem de tópicos (3) com o uso do Bertopic para gerar o dataset (compreendendo: tratamento de dados, análise quantitativa e qualitativa de dados, ajuste de dados e geração do DataSet), classificação com Zero-shot (4), treinamento supervisionado (5) e comparação de resultados (6) que será abordada nos resultados obtidos. A Figura 1 representa as etapas que foram desenvolvidas neste estudo.

Figura 1: Etapas do estudo



Fonte: Dados da pesquisa.

A. *Carga de Dados*

As amostras dos dados das bases do TCM (Tribunal de Contas do Município) e do TCE (Tribunal de Contas do Estado) foram carregados na tabela tLicitacaoDataSet, somando um total de 19.516 registros. Para a seleção desses dados, foram escolhidas todas as licitações que continham as palavras-chave relacionadas com "Assessoria Jurídica", totalizando 9.715 registros. Com o intuito de obter uma maior variedade de dados, foram incluídos aleatoriamente outros 9.801 documentos que não continham as palavras-chave associadas com "Assessoria Jurídica". É importante salientar que a princípio o objetivo é conseguir fazer a classificação das contratações dos serviços de consultoria jurídica de maneira mais eficiente que a utilização dos filtros das palavras-chave. Cada linha referente a uma licitação recebeu uma identificação única (Id), a fim de rastrear as linhas no processo de validação e testes dos algoritmos utilizados neste trabalho. A figura 2 apresenta alguns registros que representam a estrutura e dados da tabela tLicitacaoDataSet.

Figura 2: Estrutura do DataSet

	Id	DsObjetoLicitacao
1	1	CONTRATAÇÃO DE EMPRESA ESPECIALIZADA PARA O FORN...
2	2	Prestação de serviços com a contratação de 850(oitocentos e cinq...
3	3	Aquisição de Moto niveladora, zero quilômetro; características: moto...
4	4	INTEGRADOR QUIMICO
5	5	Solicitação de serviço para confecção de dois pares de placas auto...
6	6	MEMORIA FLASH (PEN DRIVE DE 16 E 32 GB)
7	7	Para fornecer água para o setor de educação municipal.
8	8	Solicitação para aquisição de materiais de consumo para suprir às n...
9	9	ABRAÇADEIRA,CABOS,DISJUNTORES,REATOR
10	10	AQUISIÇÃO DE DIVERSOS MATERIAIS DE EXPEDIENTE, PARA ...
11	11	AQUISIÇÃO DE DIVERSOS MATERIAIS DE LIMPEZA E HIGIENE ...

Fonte: Dados da pesquisa.

B. *Análise exploratória dos dados*

Nessa etapa foi efetuada uma análise exploratória do dataset para entender os dados, como: identificar os tamanhos dos textos - que descrevem os objetos das licitações - e as palavras mais comuns. As descrições dos textos dos objetos das licitações são em sua grande maioria formadas por 300 caracteres. Podendo ser considerados como textos curtos.

C. *Modelagem de tópicos com o BERTopic*

O modelo de extração de tópicos utilizado neste trabalho foi BERTopic. Essa técnica foi escolhida por considerar as características semânticas das representações vetoriais, permitindo a codificação do significado do texto e aproximando os documentos semelhantes no espaço vetorial [16]. Isso proporciona a construção do DataSet com os termos mais relevantes para as classes das licitações, que foi utilizado para execução e avaliação dos algoritmos de classificação. Esta etapa possui uma natureza iterativa, sendo formada por seis tarefas: tratamento de dados, execução do BERTopic, análise quantitativa dos tópicos gerados, análise qualitativa dos mesmos, ajuste de dados, caso necessário e por fim a geração do DataSet classificado. Os processos de tratamento de dados e do BERTopic foram executados até que as análises dos resultados indicassem o melhor agrupamento dos registros com base na coerência e na observação dos gráficos de nuvens de palavras gerados para analisar qualitativamente os termos mais representativos. O objetivo após extrair os tópicos com o BERTopic é classificar os documentos no tópico de interesse: "Assessoria Jurídica".

Alguns modelos de embeddings pré-treinados em português ou multilíngue foram utilizados para a extração de tópicos com o BERTopic nesse projeto, como: all-MiniLM-L6-v2 (rápida resposta e eficiência computacional), all-mpnet-base-v2 (combina precisão e compreensão contextual, maior consumo de recursos [23]), BERTimbau (ideal para o idioma português [24]) e roBERTA (alta precisão e robustez [25]). A escolha desses modelos foi baseada na abrangência do modelo em relação ao idioma português. Vale salientar que não existe um modelo embedding perfeito, sendo necessário avaliar e testar o que melhor se ajusta ao problema proposto [15]. Além de escolher o modelo de embeddings, é necessário configurar outros parâmetros, que o BERTopic disponibiliza. Nesse trabalho, alguns desses parâmetros foram alterados de acordo com as características do nosso dataset e um deles foi a redução de dimensionalidade, o UMAP. Por padrão, o BERTopic faz uso de modelos de redução de dimensionalidade (UMAP) [26]. No estudo, foram realizados alguns experimentos com os hiperparâmetros do UMAP e do próprio BERTopic [27]. Testou-se no experimento, as funções de redução e junção [28] dos tópicos gerados. Pois, junto aos especialistas do negócio, conseguimos identificar alguns tópicos que poderiam ser unidos num único, por serem muito similares. Também conseguimos identificar um número máximo de tópicos que conseguiu agrupar de forma satisfatória os diferentes tópicos da amostragem o que melhorou

a granularidade e a quantidade de tópicos identificados pelo modelo. O BERTopic foi executado em três versões de texto, sempre testando combinações diferentes de hiperparâmetros:

1) Texto tratado: foram retiradas as stopwords e uma lista de palavras irrelevantes para o negócio, além do processamento básico que é a retirada de símbolos, espaços, normalização dos textos para letras minúsculas, retiradas de números;
2) Texto Semi-tratado: todos os caracteres foram transformados em minúsculo e foram retirados os espaços em branco, símbolos e números. Apenas os caracteres alfanuméricos foram mantidos.
3) Texto Original: sem nenhum tratamento, ou seja, os textos foram utilizados conforme o registro na base de dados origem.

Após todos os testes serem tabelados e analisados os resultados, concluiu-se que o modelo de embedding que gerou melhores resultados foi o ALL-MNET-BASE-V2, apesar de ser 5 a 8 vezes mais lento do que os demais testados. A métrica da coerência ficou em 91% e a análise qualitativa do tópico de interesse pelos analistas de negócio concluiu que o tópico 0 (“Assessoria Jurídica”), está com boa qualidade. Reuniu 7170 documentos no tópico de interesse. A análise qualitativa foi realizada nos tópicos extraídos e verificada que os tópicos gerados pelo BERTimbau e roBERTa são muito similares e trazem agrupamentos coerentes dentro do que era esperado. Segue alguns resultados tabelados:

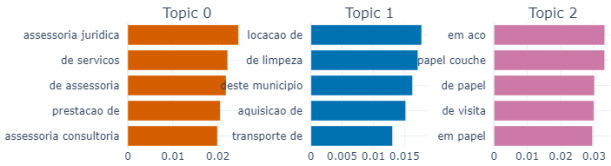
Tabela II: Resultados Modelos BerTopic

Embedding	Texto	Tópicos Gerados	Qtd Outliers	Métrica Coerência
Padrão (all-MiniLM-L6-v2)	Tratado	126	5837	73.00%
BERTimbau	Semi-tratado	59	6390	90.33%
roBERTa	Semi-tratado	59	6390	90.33%
ALL-MNET-BASE-V2	Semi-tratado	49	5103	89.77%

Fonte: Dados da pesquisa.

A tabela acima demonstra alguns resultados obtidos com as diferentes estratégias. A métrica de coerência varia bastante dependendo do hiperparâmetro e do texto utilizado pelo modelo. A terceira etapa inicia com a formação do DataSet através da definição dos grupos de documentos gerados pelo BERTopic. Os grupos que reuniram as palavras mais contextualizadas com o tema de “assessoria/consultoria jurídica” foram analisados qualitativamente através de verificações com a área de negócio e rotulados com o valor 1 (um), indicando que a classificação destes textos se refere aos serviços de consultoria jurídica. Já os outros grupos, que não pertencem a este tipo de serviço, foram classificados como 0 (zero). Segue abaixo alguns tópicos que foram agrupados pelo BERTopic.

Figura 3: Exemplos de Tópicos



Fonte: Dados da pesquisa.

O tópico 0(zero) foi o que agregou os documentos mais aderentes ao contexto de prestação de serviços de consultoria/assessoria jurídica.

D. Classificação com Zero-shot e avaliação do modelo

Na etapa de classificação não-supervisionada foram realizados os testes de classificação utilizando a técnica Zero-Shot Classification. Esta abordagem utilizou os termos mais relevantes dos tópicos levantados através do BERTopic para avaliar os resultados. Por não ser supervisionado, o modelo Zero-Shot Classification não precisou ser treinado. Como justificado acima, o 0-SHOT-TC (Zero-Shot Text Classification), se propõe a associar um rótulo apropriado a um texto, independentemente do domínio ou do aspecto (como tópico, emoção, evento, etc.). Para realizar a classificação com esta técnica, foi utilizado o pipeline de classificação zero-shot da hugging face, ZeroShotClassificationPipeline, que permite classificar documentos em tópicos pré-definidos com base nas descrições fornecidas. [29] Para realizar a classificação zero-shot em português, é essencial um modelo de linguagem que suporte o idioma citado. A Hugging Face oferece vários modelos de classificação zero-shot que são multilíngues e incluem suporte ao português, como o xlm-roberta-large-xnli [29], utilizado neste trabalho. O xlm-roberta-large-xnli é um modelo poderoso e eficiente para inferência textual multilíngue, destacando-se pela sua capacidade de generalização e desempenho superior em uma ampla gama de idiomas.[30] Testou-se duas abordagens com o Zero-Shot Text Classification:

1) Labels Candidatas: são as possíveis categorias ou rótulos que podem ser atribuídos a um texto. As labels candidatas são integradas diretamente no processo de classificação. Primeiro mapeia o texto e as labels para um espaço comum usando a Análise Semântica Explícita (Explicit Semantic Analysis - ESA) e depois escolhe a label com a maior pontuação de correspondência. Essa abordagem enfatiza que a representação das labels é tão crucial quanto a aprendizagem da representação do texto [18]. Vários testes foram realizados com a construção das Labels, e constatado a veracidade desta afirmação. A label que trouxe o melhor resultado foi “serviços advocatícios”. A partir dessa label, a hipótese foi desenvolvida: “Este texto é sobre serviços advocatícios?”.
2) Hipóteses: o uso de hipóteses envolve a construção de

uma suposição ou declaração que pode ser verificada em relação ao texto. Isso imita como os humanos decidem a veracidade das labels de qualquer aspecto. Normalmente, os humanos entendem o problema descrito pelo aspecto e o significado das labels candidatas, depois constroem mentalmente uma hipótese preenchendo uma label candidata, por exemplo, "esportes", no problema definido pelo aspecto "o texto é sobre?". Então, perguntam a si mesmos se essa hipótese é verdadeira, dado o texto [18].

Ambas as técnicas têm suas vantagens e podem ser escolhidas com base nas especificidades do problema e dos dados disponíveis. Como é feito [18]:

1) Formulação do Problema:
   a) Premissa: O texto que você deseja classificar.
   b) Hipótese: utiliza uma frase que representa uma das possíveis
2) Relações de Inferência:
   a) Para cada possível classe, gera uma hipótese e utiliza o modelo de NLI para determinar a relação entre a premissa (texto) e a hipótese (classe).
   b) A classe que resultar na maior probabilidade de implicação é considerada a classificação do texto.

Exemplo de como classificar o texto na categoria “Assessoria Jurídica”, tópico conhecido e extraído usando o BERTopic. Neste caso, só terá uma classe, mas poderia ter mais de uma, então o modelo iria escolher a que trouxesse maior probabilidade. Utiliza-se a probabilidade para dizer se é ou não, “assessoria jurídica”. Exemplificando com um dos textos:

1) Formulação do Problema:
   a) Texto: "A execução de serviços profissionais de advocacia especializada, com eventual propositura de ações judiciais de interesse do SAAE, defesa judiciais e administrativa deste, elaboração de pareceres "
   b) Hipótese: “Este texto é sobre serviços advocatícios?”.
2) Relações de Inferência:
   a) Uso do Modelo de NLI: a premissa (texto) e a hipótese são passadas pelo modelo de NLI. O modelo avalia a relação entre o texto e a hipótese.

Classificação: Quando houver mais de uma hipótese, a com a maior pontuação de "implicação"ou menor pontuação de "contradição"determina a categoria do texto. Neste caso, se a probabilidade de “verdadeira” for alta o suficiente (por exemplo, >0.8), o texto é classificado como relevante para a hipótese, e a classificação é true para Assessoria Jurídica. O resultado conterá as seguintes informações:

Tabela III: Exemplo Classificação Label Candidata Zero-Shot

Labels	*Scores*
Este texto é sobre serviços advocatícios	0.99

Fonte: Dados da pesquisa.

Sendo que Labels significa a lista de hipóteses fornecidas e o Scores as probabilidades associadas a cada classe, indicando a confiança do modelo de que a hipótese é verdadeira em relação ao texto. Com as Labels Candidatas, o processo é muito similar com a presença da premissa e uma lista de possíveis labels que podem ser atribuídos a um texto substituindo as hipóteses. A relação de Inferência o modelo de NLI determinará a relação entre a premissa(texto) e label(classe). A label com maior pontuação de correspondência será considerada a classificação do texto. Na análise com as áreas de negócios foi avaliado que a acurácia e a precisão são as métricas mais aderentes para a solução do problema, pois a acurácia informa o quanto as licitações foram classificadas corretamente, independentemente da classe. Já a Precisão enfatiza o quanto o modelo acerta quando classifica um tipo como positivo. Seguem as métricas obtidas na classificação com o Zero-Shot Classification:

Tabela IV: Métrica do Zero-Shot informando um tópico

Texto Tratado

Tipo	Valor
Acurácia	87%
Precisão	88%

Texto semi-tratado

Tipo	Valor
Acurácia	88%
Precisão	88%

Texto original

Tipo	Valor
Acurácia	88%
Precisão	90%

Fonte: Dados da pesquisa.

Tabela V: Métrica do Zero-Shot utilizando hipótese

Texto Tratado

Tipo	Valor
Acurácia	82%
Precisão	88%

Texto semi-tratado

Tipo	Valor
Acurácia	84%
Precisão	88%

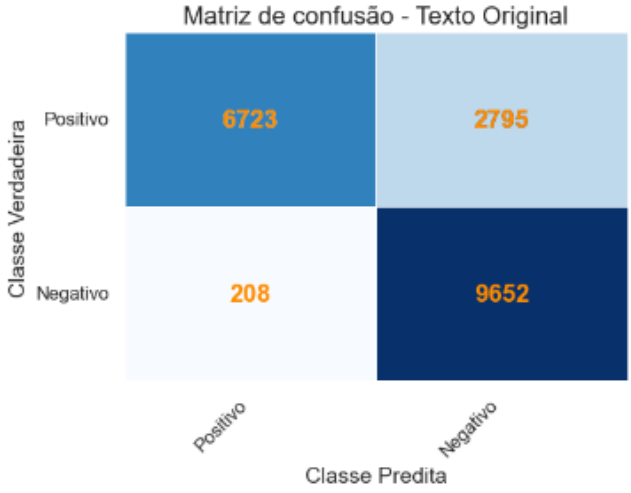
Texto original

Tipo	Valor
Acurácia	83%
Precisão	90%

Fonte: Dados da pesquisa.

Apesar da acurácia ter ficado um pouco melhor no texto semi-tratado, o texto original nos traz maior precisão, ou seja, 90% dos textos originais que foram classificados como “Assessoria jurídica” pertencem verdadeiramente a esta classe, segundo a classificação do BERTopic. Deve ficar claro aqui que o BERTopic, pode ter classificado errado. A seguir, a figura 4, mostra a matriz de confusão do Zero-Shot aplicado no Texto Original, usando a abordagem de label candidata.

Figura 4: Matriz de confusão – Zero-Shot usando label candidata



E. Classificação com Random Forest e sua avaliação

Após o treinamento não-supervisionado foi realizado o treinamento supervisionado através do uso da biblioteca PyCaret, utilizando o mesmo dataset gerado pela etapa do BERTopic. No processo de treinamento os documentos foram submetidos ao tratamento com a normalização dos textos para caixa baixa, remoção das pontuações, stopwords, caracteres especiais e números. Após o tratamento de dados, os textos foram submetidos ao PyCaret. A tabela de ranqueamento de modelos do PyCaret apresentou os seguintes resultados:

Tabela VI: Análise Modelos PyCaret

Modelo	Acurácia	Precisão	Recall	F1
Random Forest Classifier	0.9013	0.8017	0.935	0.8631
Decision Tree Classifier	0.6934	0.6367	0.6451	0.6407
Support Vector Machine	0.5284	0.4697	0.5330	0.4991
Logistic Regression	0.4329	0.4135	0.3911	0.3373

Fonte: Dados da pesquisa.

Após a análise da tabela, o modelo Random Forest apresentou a melhor precisão e acurácia com valores respectivos de 80,17% e 90,13%. Diante destes resultados o modelo foi tunado com os seguintes parâmetros: n_estimators = 100, criterion = “gini”, max_depth=None, min_samples_split=2,min_samples_leaf=1, min_weight_fraction_leaf==0.0, max_features=’sqrt’, max_leaf_nodes=None, min_impurity_decrease=0.0, random_state=42, bootstrap=True, oob_score=False, n_jobs=-1, verbose=0, warm_start=False, max_samples=None. Com essa configuração o modelo obteve os seguintes resultados com os dados de treinamento:

Tabela VII: Resultados Random Forest Classifier

Modelo	Acurácia	Precisão	Recall	F1
Random Forest Classifier	0.86	0.93	0.87	0.89

Fonte: Dados da pesquisa.

IV Conclusão

Com o objetivo de apoiar a análise de aquisições do estado da Bahia através do BI Licitômetro, agrupando as licitações através da descrição dos objetos dos processos de compra, este estudo comparou modelo supervisionado com modelo não supervisionado para classificar as licitações referentes aos serviços de assessoria/consultoria jurídica. Neste trabalho foram aplicados a modelagem de tópicos e a classificação das licitações referentes ao tema "Assessoria/Consultoria Jurídica". Os

dados foram coletados nas bases do TCE e TCM, os modelos de embeddings BERTimbau, RoBERTa, All-MiniLM-L6-v2 e All-mpnet-base-v2 foram testados na modelagem de tópicos. O modelo de classificação não-supervisionado Zero-Shot foi aplicado assim como o modelo supervisionado Random Forrest Classifier. Os dois modelos apresentaram bom desempenho, porém o Zero-Shot foi mais eficaz. A execução dos métodos alcançou os objetivos de rotular e classificar as contratações de serviços advocatícios pelos órgãos públicos do estado da Bahia dentro do contexto avaliado. Além disso, foi observado que o Zero-Shot foi capaz de interpretar corretamente contextos mais variados que o BERTopic e o Random Forest conforme os tópicos listados abaixo.

1) LOCAÇÃO DE IMÓVEL CASA PARA O FUNCIONAMENTO DA ASSESSORIA JURÍDICA DESTE MUNICIPIO DE CAETITE-BAHIA.:
   a) Palavras-Chave
      - Classificação Assessoria Jurídica: 1
   b) BERTopic
      - Classificação Assessoria Jurídica: 1
   c) Zero-Shot
      - Classificação Assessoria Jurídica: 0
   d) Classificação Random Forest
      - Classificação Assessoria Jurídica: 0
   e) Valor Esperado
      - Classificação Assessoria Jurídica: 0
2) REFERENTE A PRESTAÇÃO DE SERVIÇOS ESPECIALIZADOS DE ASSESSORIA E CONSULTORIA JURÍDICA AO MUNICIPIO DE ITAPEBI-BA, CONF. CONTRATO 002/2018:
   a) Palavras-Chave
      - Classificação Assessoria Jurídica: 1
   b) BERTopic
      - Classificação Assessoria Jurídica: 1
   c) Zero-Shot
      - Classificação Assessoria Jurídica: 1
   d) Classificação Random Forest
      - Classificação Assessoria Jurídica: 0
   e) Valor Esperado
      - Classificação Assessoria Jurídica: 1
3) FORNECIMENTO DE LANCHES DESTINADAS A ASSESSORIA JURÍDICA, ASSESSORIA CONTÁBIL E SERVIDORES DESTA CÂMARA.:
   a) Palavras-Chave
      - Classificação Assessoria Jurídica: 1
   b) BERTopic
      - Classificação Assessoria Jurídica: 1
   c) Zero-Shot
      - Classificação Assessoria Jurídica: 0
   d) Classificação Random Forest
      - Classificação Assessoria Jurídica: 1
   e) Valor Esperado
      - Classificação Assessoria Jurídica: 0
4) ATENDER DESPESA REFERENTE A DIÁRIAS DESTINDAS A COBRIR DESPESAS COM ALIMENTAÇÃO E HOSPEDAGEM PARA PARTICIPAR DE REUNIÃO NA SUDESB E REUNIÃO COM OS ADVOGADOS, NA CIDADE DE SALVADOR:
   a) Palavras-Chave
      - Classificação Assessoria Jurídica: 1
   b) BERTopic
      - Classificação Assessoria Jurídica: 0
   c) Zero-Shot
      - Classificação Assessoria Jurídica: 0
   d) Classificação Random Forest
      - Classificação Assessoria Jurídica: 0
   e) Valor Esperado
      - Classificação Assessoria Jurídica: 0
5) DESPESA COM AQUISIÇÃO DE MATERIAL DE CONSTRUÇÃO (REJUNTE, PISO E ARGAMASSA) PARA TROCA DO PISO DA SALA DA ASSESSORIA JURÍDICA, VISANDO ATENDER AS NECESSIDADES DA CÂMARA MUNICIPAL.
   a) Palavras-Chave
      - Classificação Assessoria Jurídica: 1
   b) BERTopic
      - Classificação Assessoria Jurídica: 1
   c) Zero-Shot
      - Classificação Assessoria Jurídica: 0
   d) Classificação Random Forest
      - Classificação Assessoria Jurídica: 0
   e) Valor Esperado
      - Classificação Assessoria Jurídica: 0

Nos textos de exemplo, podemos observar que usando as palavras-chave, todos os textos foram classificados como 1(Assessoria/Consultoria Jurídica), pois contém as palavras: Assessoria Jurídica, Consultoria Jurídica, Advogados. Analisando o valor esperado com os resultados do BERTopic, Random Forest e Zero Shot, concluímos que o Zero Shot Classification é capaz de identificar o contexto correto, se mostrando mais adequado para possíveis variações de descrições de objetos de licitações do tipo Assessoria Jurídica.

V Trabalhos Futuros

Como trabalhos futuros, é possível citar a necessidade de classificar outros temas relevantes para rotulação e classificação das licitações coletadas nas bases. Esta classificação agregará mais valor às consultas no Licitômetro. Assim como, possibilitará o desenvolvimento de outros serviços de IA para apoio na fiscalização das Licitações.

Referências

[1] V. A. J. d. Amorim, "Licitações e contratos administrativos : teoria e jurisprudência," 2020, acessado em: 08 jun. 2024. [Online]. Available: https://www2.senado.leg.br/bdsf/bitstream/handle/id/573456/licitacoes_contratos_administrativos_3ed.pdf

[2] C. Nacional, "Lei de licitações e contratos administrativos," 2024, acessado em: 08 jun. 2024. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm

[3] ——, "Lei nº 12.846, de 1º de agosto de 2013," 2024, acessado em: 08 jun. 2024. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2013/lei/l12846.htm

[4] B. V. S. Vinces, "Análise em larga escala da evolução temporal de tópicos obtidos do twitter basado em apache spark," 2022, acessado em: 15 jun. 2024. [Online]. Available: https://repositorio.usp.br/directbitstream/9d6800b1-3f33-4a6c-94b0-eb15f02ab7a7/Braulio%20Valentin%20S%C3%A1nchez%20Vinces_Monograf%C3%Ada%20vers%C3%A3o%20final%20-%20Projeto%20de%20Pesquisa%20-%20Braulio%20Valentin%20S%C3%A1nchez%20Vinces_206502.pdf

[5] N. S. Costa, "Aplicação de técnicas de inteligência artificial para análise de tweets sobre institutos federais brasileiros," https://repositorio.ifg.edu.br/bitstream/prefix/1932/1/TCC%20N%C3%A1tallya%20Soares.pdf, 2023, acessado em: junho de 2024.

[6] F. A. B. Érika Kayoko Hamaguti, "Introdução sobre machine learning e deep learning," 2022, acessado em: junho de 2024. [Online]. Available: https://www.fabriciobreve.com/artigos/2022/jornacitec2022expandido.pdf

[7] B. Mondal, Artificial Intelligence: State of the Art, January 2020, pp. 389–425.

[8] T. Mitchell, Machine Learning. Boston, MA: WCB / McGraw-Hill – Computer Science Series, 1997.

[9] G. M. d. M. Paixão, B. C. Santos, R. M. d. Araujo, M. H. Ribeiro, J. L. d. Moraes, and A. L. Ribeiro, "Machine learning na medicina: Revisão e aplicabilidade," Arquivos Brasileiros de Cardiologia, vol. 118, no. 1, p. 95–102, Jan 2022. [Online]. Available: https://doi.org/10.36660/abc.20200596

[10] T. B. Ludermir, "Inteligência artificial e aprendizado de máquina: estado atual e tendências," Estudos Avançados, vol. 35, no. 101, p. 85–94, Jan 2021. [Online]. Available: https://doi.org/10.1590/s0103-4014.2021.35101.007

[11] G. G. Chowdhury, "Natural language processing," Annual Review of Information Science and Technology, vol. 37, no. 1, pp. 51–89, 2003. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103

[12] V. J. Alles, "Construção de um corpus para extrair entidades nomeadas do diário oficial da união utilizando aprendizado supervisionado," 2019, acessado em: junho de 2024. [Online]. Available: http://www.realp.unb.br/jspui/bitstream/10482/34901/1/2018_VanderleiJandirAlles.pdf

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[14] I. C. P. Albineli, "Análise de sentimentos de publicações em plataformas on-line sobre turismo no brasil," 2023, acessado em: junho de 2024. [Online]. Available: https://adelpha-api.mackenzie.br/server/api/core/bitstreams/1f4e4cc6-f224-486f-a122-0ad48ba12944/content

[15] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022. [Online]. Available: https://arxiv.org/abs/2203.05794

[16] G. B. B. Gomes, "Contribuições à análise histórica e social em rede social baseada em processamento de linguagem natural," https://repositorio.unicamp.br/Busca/Download?codigoArquivo=565542&tipoMidia=0, 2023, acessado em: junho de 2024.

[17] B. M. Gêda, "Classificação de textos de decisões judiciais," https://www.repositorio.ufal.br/bitstream/123456789/10441/1/Classifica%C3%A7%C3%A3o%20de%20textos%20de%20decis%C3%B5es%20judiciais.pdf, 2021, acessado em: junho de 2024.

[18] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," 2019. [Online]. Available: https://arxiv.org/abs/1909.00161

[19] A. I. Navarro, "Zero-shot learning in nlp," 2024, acessado em: Agosto de 2024. [Online]. Available: https://modulai.io/blog/zero-shot-learning-in-nlp/

[20] E. S. e Manoel Villas Bôas Júnior e Wagner Luiz Lobo Ferreira, "A importÂncia de utilizar mÉtricas adequadas de avaliaÇÃo de performance em modelos preditivos de machine learning," Revista Projectus Rio de Janeiro, vol. 7, no. 2, pp. 52–62, 2022, acessado em: julho de 2024. [Online]. Available: https://doi.org/10.15202/25254146.2022v7n2p52

[21] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ser. WSDM '15. New York, NY, USA: Association for Computing

Machinery, 2015, p. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324

[22] L. Rocha, D. Welter, and D. Duarte, “Modelagem probabilística de tópicos: Uma comparação empírica,” in Anais da XVI Escola Regional de Banco de Dados. Porto Alegre, RS, Brasil: SBC, 2021, pp. 41–50. [Online]. Available: https://sol.sbc.org.br/index.php/erbd/article/view/17237

[23] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “Mpnet: Masked and permuted pre-training for language understanding,” 2020. [Online]. Available: https://arxiv.org/abs/2004.09297

[24] F. Souza, R. Nogueira, and R. Lotufo, “BERTimbau: pretrained BERT models for Brazilian Portuguese,” in 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear), 2020.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[26] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[27] T. L. e Márcia de Lima, “Pln – processamento de linguagem natural para iniciantes,” 2021, acessado em: Agosto de 2024. [Online]. Available: https://www.insightlab.ufc.br/

[28] M. Grootendorst, “Bertopic,” 2024, acessado em: Agosto de 2024. [Online]. Available: https://maartengr.github.io/BERTopic/

[29] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” 2020. [Online]. Available: https://arxiv.org/abs/1911.02116

[30] N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau, “Larger-scale transformers for multilingual masked language modeling,” 2021. [Online]. Available: https://arxiv.org/abs/2105.00572

Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

CENTRO UNIVERSITÁRIO SENAI CIMATEC
ESPECIALIZAÇÃO EM DATA SCIENCE & ANALYTICS

ATA DE APRESENTAÇÃO DE PROJETO FINAL DE CURSO

Ata de apresentação do Projeto Final de Curso, "**Descoberta de tópicos e classificação de textos de licitações promovidas pelos órgãos públicos do estado da Bahia"**, submetido pela aluna Rubia Teles de Souza, como parte dos requisitos para obtenção do Certificado de **Especialista em Data Science & Analytics** pelo Centro Universitário SENAI CIMATEC, às 20h00 do dia 27 de Agosto de 2024. Reuniu-se remotamente pela plataforma Google Meet, a Banca Examinadora designada pelo Prof MSc Braian Varjão Gama Bispo – Orientador, constituída pelo Prof MSc Braian Varjão Gama Bispo e Prof Dr. Lilian Cristina da Silveira. O Orientador deu início aos trabalhos com as devidas orientações, e a exposição foi realizada pelo estudante dentro do prazo de tempo estabelecido. Ao final da apresentação a banca reuniu-se atribuindo a seguinte nota: **9,1** (nove pontos e um décimo).

A banca de avaliadores decidiu pela:

(X) Aprovação do trabalho
Caberá ao aluno apresentar em no máximo em 30 (trinta) dias a contar da data de assinatura desta Ata, uma cópia do trabalho em PDF, constando as considerações pontuadas pela banca. A Ata de Apresentação do Projeto Final de Curso deve ser digitalizada e inserida na terceira página do TCC ou como anexo do artigo.

( ) Reprovação do trabalho
O aluno terá que se matricular novamente no TCC – Trabalho de Conclusão de Curso e ser submetido a uma banca avaliadora no semestre seguinte**.**

As ações consequentes ao status de Aprovação deverão obedecer ao prazo proposto acima sob pena do parecer final ser modificado para o status de Reprovado automaticamente e sem possibilidade de recurso.

Para constar, lavrou-se a presente ata que vai assinada por todos os membros da Banca. Por estarem cientes de suas obrigações estão de acordo com os termos desse documento:

Salvador, 27 de Agosto de 2024.

gov.br Documento assinado digitalmente
BRAIAN VARJAO GAMA BISPO
Data: 31/08/2024 00:37:10-0300
Verifique em https://validar.iti.gov.br

Braian Varjão Gama Bispo
Professor Orientador

gov.br Documento assinado digitalmente
LILIAN CRISTINA DA SILVEIRA
Data: 02/09/2024 20:37:04-0300
Verifique em https://validar.iti.gov.br

Lilian Cristina da Silveira
Membro da banca

Assinado eletronicamente por:
Patricia Freitas Tourinho
CPF: ***.733.265-**
Data: 05/09/2024 17:55:09 -03:00