Segmentation and Summarization for Extracting Information about Information Technology Equipment from Government Procurement Notice

Erick Correia Silva[1], Ivo Paixão de Medeiros[2], Maria Viviane de Menezes[1], Dayse Simon Landim Kamikawachi[2]

[1] Programa de Pós-Graduação em Computação (PCOMP)
Universidade Federal do Ceará - Campus Quixadá
Quixadá, Ceará, Brasil
erickbastos.cs@alu.ufc.br, vivianemenezes@ufc.br
[2] Centro de Excelência em Inteligência Artificial (CEIA)
Universidade Federal do Goiás
Goiânia, Goiás , Brasil
ivopdm@gmail.com, daysesimon@gmail.com

Abstract. Government procurement in Brazil employs a bidding process to acquire products and services, involving stages such as the publication of public notices, which are structured documents outlining procurement rules and specifications. For Information Technology (IT) companies, competitive participation in the bidding process includes monitoring opportunities by analyzing data from these notices. This paper applies text segmentation and summarization algorithms to extract data such as product names, prices and quantities from IT procurement notices. Four architectures are proposed: (i) sentence-based segmentation followed by K-means clustering; (ii) section-based segmentation followed by K-means clustering; (iii) sentence-based segmentation followed by BERTimbau clustering; and (iv) section-based segmentation followed by BERTimbau clustering. For all texts clustered as an interest profile, the Large Language Model (LLM) GPT-3.5 is applied in order to summarize and organize the information regarding product names, prices and quantities. Evaluation using real public notices from Federal and State Government Procurement sites shows that BERTimbau significantly outperformed K-means in both sentence and section segmentation tasks.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: Segmentation, Summarization, Government Procurement

1. INTRODUÇÃO

Licitação é um procedimento utilizado pela administração pública brasileira para adquirir produtos, obras, serviços e alienações [da União 2024]. Regulada pela Lei nº 14.133/2021 [da República 2021], a licitação compreende sete fases: preparatória, divulgação do **edital de licitação**, apresentação de propostas e lances (quando aplicável), julgamento, habilitação, recursal e homologação [dos Santos Chaves 2015]. A fase preparatória envolve a elaboração dos documentos e estudos necessários. Na divulgação do edital, o documento contendo regras e especificações é publicado, dividido em seções como objeto da licitação, condições de participação e critérios de julgamento. Na apresentação de propostas e lances, empresas interessadas submetem suas propostas conforme o edital. O julgamento avalia as propostas, seguido pela habilitação, que verifica a capacidade técnica e jurídica dos participantes. A fase recursal permite recursos contra decisões, garantindo transparência e equidade. A homologação ratifica o resultado, autorizando a formalização do contrato-

Para empresas que vendem equipamentos tecnológicos, a participação competitiva em licitações requer a identificação de oportunidades de vendas nos editais que estão disponíveis, geralmente em arquivos no formato PDF, em sites dos governos federais, estaduais e municipais. No entanto, os editais são documentos longos, muitas vezes trazendo dados não-estruturados, e, assim, tarefas de Processamento de Linguagem Natural (PLN) [Jurafsky and Martin 2023], tais como segmentação e sumarização [da Silva et al. 2022; Cho et al. 2022], podem ajudar na extração automática de informações relevantes desses documentos. Segmentação é a tarefa de dividir um documento em partes menores (como sentenças ou seções) que podem ser tratadas de forma independente. Já a sumarização é o processo de criar uma versão concisa de um texto que capture suas ideias principais.

Este trabalho utiliza segmentação e sumarização de textos, visando a extração de informações sobre produtos, valores e quantidades em editais de licitação de produtos tecnológicos. Propomos quatro arquiteturas distintas para essa tarefa: (i) a primeira, usa segmentação por sentenças, seguida de clusterização com o algoritmo *K-means* [Ahmed et al. 2020]; (ii) a segunda, usa segmentação por seções, seguida de clusterização com *K-Means*; (iii) a terceira, usa segmentação por sentenças, seguida de clusterização com *BERTimbaum* [Souza et al. 2020] e; finalmente, (iv) a quarta, usa segmentação por seções, seguida de clusterização com *BERTimbaum*. Todos os textos clusterizados considerados de acordo com o perfil de interesse são posteriormente enviados para serem sumarizados pelo *Large Language Model* GPT (*Generative Pre-trained Transformer*) 3.5 [OpenAI 2023], obtendo assim as informações referentes a produtos, valores e quantidades de produtos tecnológicos contidas no edital avaliado. As arquiteturas foram validadas com editais reais obtidos dos sites de compras públicas de Governos Federal e Estaduais. O restante deste artigo está organizado como a seguir: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute trabalhos relacionados; a Seção 4 descreve a metodologia; a Seção 5 aborda os resultados obtidos; e a Seção 6 traz as conclusões e sugestões para trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Essa seção apresenta a fundamentação teórica do trabalho. Na Seção 2.1, abordamos brevemente os principais conceitos de aprendizado de máquina. Em seguida, na Seção 2.2, são discutidas as principais arquiteturas de redes neurais. As Seções 2.3 e 2.4 exploram, respectivamente, os conceitos de segmentação e sumarização de documentos.

2.1 Aprendizado de Máquina

Aprendizado de Máquina [Russell and Norvig 2016] é a subárea da inteligência artificial que se concentra no desenvolvimento de modelos e algoritmos que permitem ao agente inteligente aprender a partir de dados. Nesta área, há três formas principais para que o agente aprenda: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço.

No aprendizado não supervisionado, o agente aprende padrões mesmo sem receber um *feedback* explícito. A tarefa mais comum de aprendizado não supervisionado é a clusterização: detectar *clusters* potencialmente úteis de exemplos de entrada. Um algoritmo bastante utilizado para tarefas de clusterização é o *K-means* [Ahmed et al. 2020] que agrupa textos baseando-se em semelhanças semânticas para facilitar a identificação de padrões ou temas recorrentes. No aprendizado por reforço, o agente aprende a partir de uma série de reforços (recompensas ou punições). Um algoritmo clássico de aprendizado por reforço é o *Q-Learning* [Watkins and Dayan 1992]. No aprendizado supervisionado, o agente observa alguns pares de entrada-saída de exemplo e aprende uma função que mapeia a entrada para a saída [Russell and Norvig 2016]. Nesta categoria há os algoritmos de regressão linear [Su et al. 2012], regressão logística [LaValley 2008], árvores de decisão [Somvanshi et al. 2016], dentre outros.

2.2 Redes Neurais Artificiais

As *Redes Neurais Artificiais* (RNAs) [McCulloch 1943] são modelos computacionais inspirados no cérebro humano. Elas consistem em camadas de neurônios conectados, onde cada neurônio recebe entradas, aplica uma função de ativação e gera uma saída. As informações são processadas pelas camadas até a geração da saída. A arquitetura da rede depende do tipo de problema a ser resolvido [Russell and Norvig 2016]. Exemplos incluem *Multilayer Perceptrons* (MLPs), CNNs e RNNs. As RNAs são aplicáveis em aprendizado supervisionado, não supervisionado e por reforço.

As Redes Neurais Transformacionais (*Transformers*) representaram uma evolução significativa na área de redes neurais, especialmente para tarefas relacionadas ao PLN. Desenvolvidos inicialmente por [Vaswani et al. 2023], os *Transformers* eliminam a necessidade de recorrência nas redes neurais, utilizando mecanismos de atenção para capturar dependências globais entre as entradas e saídas. Isso permite que o modelo processe dados em paralelo e capture contextos complexos mais eficientemente. Essa arquitetura tem sido a base para modelos como o BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2019] e o GPT [Yenduri et al. 2024].

BERT [Devlin et al. 2019] utiliza a arquitetura dos *Transformers* na tarefa de compreensão de texto de forma bidirecional, o que significa que ele é treinado para entender o contexto de ambos os lados de cada palavra simultaneamente. Durante o treinamento, o BERT usa um processo chamado *Masked Language Model* (MLM), no qual algumas palavras são aleatoriamente ocultadas e o modelo aprende a prever essas palavras com base no contexto fornecido pelas palavras não ocultadas de ambos os lados. Esse método permite que cada palavra seja representada de maneira mais rica e contextualizada, incorporando informações de todo o texto. O BERT inspirou modelos derivados e aprimorados que utilizam sua arquitetura básica e a expandem como o RoBERTa (*Robustly Optimized BERT Pretraining Approach*) [Liu et al. 2019], DeBerta (*DeBERTa: Decoding-enhanced BERT with Disentangled Attention*) [He et al. 2020] e o Bertimbau [Souza et al. 2020].

O modelo BERTimbau [Souza et al. 2020] foi treinado com o corpus BrWaC (*Brazilian Web as Corpus*), um dos maiores e mais abrangentes *corpus* disponíveis para o português brasileiro, contendo aproximadamente 2,7 bilhões de palavras coletadas de páginas da *web*. Os resultados do BERTimbau demonstraram sua superioridade em várias tarefas de PLN específicas para o português brasileiro. O nome do modelo uma junção das palavras "BERT" com "berimbau".

2.3 Segmentação de Documentos

A segmentação de documentos tem como objetivo dividir um documento em partes menores, denominadas segmentos ou *tokens*, que podem ser tratadas de forma independente [da Silva et al. 2022]. Esta tarefa é essencial para a análise detalhada e estruturada de textos longos. A segmentação pode ser realizada com base em diferentes critérios, como: tópicos, frases ou entidades nomeadas. Na segmentação por tópico, o documento é dividido em segmentos baseados em mudanças de tópico ao longo do texto [Hearst 1997]. Na segmentação por frase, o texto é dividido em unidades menores que compartilham uma coesão semântica [Glavaš et al. 2016]. Já na segmentação como reconhecimento de entidades nomeadas (NER), os segmentos são considerados entidades e o modelo deve identificá-los e classificá-los [da Silva et al. 2022]. Cada uma dessas abordagens tem suas vantagens e desafios e a escolha do algoritmo de aprendizagem depende do tipo de documento, da granularidade desejada da segmentação e da disponibilidade de dados anotados para treinamento e avaliação dos modelos.

2.4 Sumarização de Documentos

A sumarização de documentos [Cho et al. 2022] é o processo de criar uma versão concisa e coerente de um documento extenso que capture suas ideias principais. Esta tarefa é fundamental para facilitar a compreensão e a extração de informações importantes de textos longos. A sumarização pode ser

realizada por meio de duas abordagens principais: extrativa e abstrativa. A sumarização extrativa envolve a seleção de sentenças ou frases significativas diretamente do texto original, combinando-as para formar um resumo que preserva a integridade do conteúdo original sem reformulação. Por outro lado, a sumarização abstrativa busca gerar um resumo reescrevendo e reestruturando as informações, criando novas sentenças que não necessariamente estão presentes no texto original, mas que capturam o seu significado essencial de forma mais condensada [Awasthi et al. 2021].

3. TRABALHOS RELACIONADOS

O artigo [Cho et al. 2022] explora a relevância da segmentação de texto para a sumarização de documentos longos. Os autores propõem uma abordagem que segmenta e sumariza seções simultaneamente, usando representações de sentenças. Eles aplicam expressões regulares para selecionar sentenças diversas, e o modelo, chamado *Lodoss*", utiliza a arquitetura *Longformer*, incorporando uma matriz de *embeddings* posicionais para captar o contexto local e global das seções. O modelo é treinado para segmentar seções e extrair sentenças salientes, utilizando o regularizador DPP, que garante sentenças informativas e diversificadas. Os experimentos mostram que o *Lodoss*" tem bom desempenho na sumarização extrativa e abstrativa, em precisão e diversidade. Análises confirmam que a segmentação melhora a qualidade dos resumos. Nossa abordagem difere ao separar os processos de segmentação e sumarização, focando exclusivamente em sumarização extrativa com o modelo GPT-3.5.

No trabalho [da Silva et al. 2022], os autores tratam da segmentação e anotação de documentos usando Reconhecimento de Entidades Nomeadas (NER) em textos do Diário Oficial do Distrito Federal. O estudo segmenta e classifica textos utilizando modelos baseados em CRF, CNN, LSTM e biLSTM-CRF, com avaliação em segmentos de palavras e frases, destacando o CRF baseado em frases como o mais eficaz. A principal contribuição é um conjunto de 127 textos anotados manualmente, em um domínio com dados menores e anotados automaticamente. Nosso trabalho é similar ao de [da Silva et al. 2022], pois também lida com textos governamentais, porém focamos em editais de produtos tecnológicos, enquanto [da Silva et al. 2022] trata de diários oficiais. Além disso, utilizamos segmentação por sentença e seção, enquanto eles empregam palavras e sentenças. Outra diferença é que nosso trabalho envolve sumarização, ausente em [da Silva et al. 2022].

O trabalho [ANDRADE and BAPTISTA 2022], por sua vez, visa automatizar a análise e auditoria de **documentos de licitação** dispostos em documentos em formato PDF. Os autores desenvolveram um modelo de aprendizagem supervisionado capaz de identificar informações específicas contidas em **editais de licitação**. O objetivo é checar o documento e seus anexos, garantindo que todas as informações necessárias estejam presentes e não haja inconsistências. Os autores comentam sobre a complexidade desta tarefa devido à natureza não estruturada dos documentos, que frequentemente incluem imagens e tabelas que dificultam a extração eficiente das informações. Para a implementação, foram utilizadas bases de dados extraídas do Portal do Governo do Estado do Acre. A metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Schröer et al. 2021] guiou o desenvolvimento, passando por fases como entendimento do problema, entendimento e preparação dos dados, modelagem e avaliação. Diferentes algoritmos de aprendizado de máquina, incluindo árvores de decisão, *Support Vector Machine* (SVM) e BERTimbau foram testados. O BERTimbau obteve os melhores resultados, destacando-se na tarefa de classificação de sentenças dos documentos em classes predefinidas. O trabalho de [ANDRADE and BAPTISTA 2022] assemelha-se a nosso, pois tem o objetivo de obter informações contidas em editais de licitação, além de usar modelos de classificação de sentenças como o BERTimbau. As diferenças em relação ao nosso trabalho é que usamos também algoritmos de aprendizagem não supervisionados para a terefa de segmentação e o LLM GPT 3.5 para a tarefa de sumarização.

4. METODOLOGIA

O objetivo deste artigo é propor uma abordagem para segmentação e sumarização de textos de editais de licitação de **produtos tecnológicos** com o objetivo de **extrair informações sobre produtos, quantidades e valores**. A Seção 4.1 descreve o conjunto de dados utilizado neste trabalho. A Seção 4.2 apresenta as duas arquiteturas propostas para realização desta tarefa.

4.1 Conjunto de Dados

O conjunto de dados é um conjunto de editais de licitações de aquisição de produtos tecnológicos, dispostos no formato PDF. De forma geral, os editais possuem informações como: (i) os requisitos da licitação, (ii) os critérios de elegibilidade, (iii) definição e descrição dos objetos e (iv) prazos e procedimentos necessários para a submissão de propostas. Mesmo que haja uma legislação que regulamente as informações que devam ser disponibilizadas nos editais, a forma de apresentação pode se diferenciar de um documento para outro, a depender do órgão contratante. Nesta pesquisa, consideramos editais dispostos em um arquivo único. Em grande parte, os arquivos foram disponibilizados em formato PDF. Apenas uma pequena porção (menos de 5%) está em arquivo doc ou docx. O conjunto dos dados é formado por cerca de 800 editais de aquisição de produtos tecnológicos publicados em 2023 por várias esferas governamentais, com documentos variando de 1 a 150 páginas. Há desafios na microestrutura do documento, especificamente na seção que define e descreve os objetos da licitação, relacionados à identificação e extração de elementos como *identificador do item*, *identificador do lote*, *nome do produto*, *quantidade* e *especificação técnica*. Esses elementos textuais podem apresentar-se de diferentes formas, como tabelas, listas itemizadas, quadros, entre outros. Há também editais que indicam cotas reservadas sobre o tipo de empresa que pode oferecer determinados itens (e.x: apenas microempresa).

4.2 Arquitetura do Sistema

As arquiteturas propostas para extração de informação de editais de licitação estão ilustrada na Figura 1, As arquiteturas diferem principalmente na forma de segmentar o texto (passo 2) e no modelo para encontrar textos relevantes (passo 5).

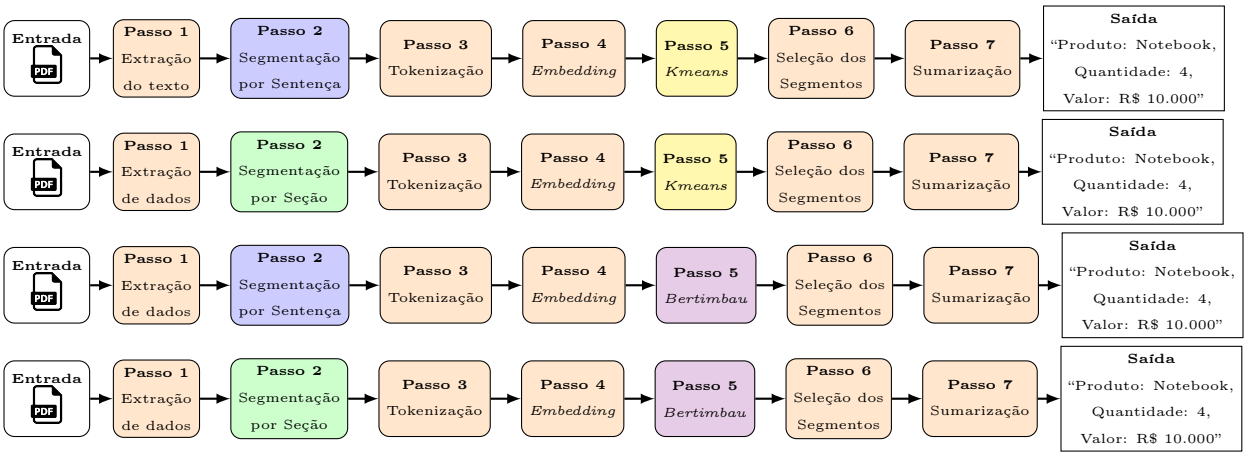

Fig. 1. Arquiteturas 1, 2, 3 e 4 apresentadas de maneira sequencial.

—**Passo 1: Extração do Texto do PDF**. O arquivo de entrada é um PDF correspondente a um edital de licitação de produtos tecnológicos. Neste passo, é realizada a extração do texto

do PDF do edital de licitação com aplicação de técnicas de processamento de texto para remover qualquer elemento indesejado que possa interferir na análise subsequente, como cabeçalhos, rodapés, numeração de páginas ou marcas d'água.

—**Passo 2: Aplicação de Algoritmos de Segmentação**. No segundo passo, o texto extraído é segmentado de duas formas: por sentenças (arquiteturas 1 e 3) e por seções (arquiteturas 2 e 4). Na segmentação por sentenças, o texto é dividido em sentenças sempre que um ponto seguido de espaço é encontrado. Na segmentação por seções o texto é dividido em suas partes estruturais (como introdução e requisitos) usando expressões regulares para identificar e separar os segmentos.

—**Passo 3: Tokenização dos segmentos de texto**. Neste passo, a sentença de texto é submetida a uma função de tokenização para determinar seu número total de *tokens*. Se o total de *tokens* for igual ou inferior ao limite estabelecido pelo *BERTimbau*, a sentença pode ser processada diretamente. Caso contrário, se o número de *tokens* exceder esse limite, será necessário dividir a sentença em partes menores. Cada segmento resultante contém um número de *tokens* que não ultrapassa o limite máximo de 512 *tokens* .

—**Passo 4: Criação das *Embeddings*.** Cada *token* é então transformado em *embeddings* usando o modelo BERTimbau.

—**Passo 5: Aplicação dos modelos de aprendizagem de máquina** Nesta etapa os *embeddings* são aplicados em dois modelos distintos de aprendizado de máquina. O primeiro é o *KMeans*. O segundo é um classificador baseado no *Bertimbau*, que usa os *embeddings* para determinar a categoria ou classificação dos segmentos de texto. As arquiteturas propostas diferem principalmente neste passo.

  —**Arquitetura 1: Segmentação por Sentença + Classificação KMeans**: Os textos segmentados por sentença têm suas embeddings agrupadas em clusters para identificar qual cluster contém o maior número de textos de interesse. Dessa forma, quando uma nova instância for criada, é possível identificar sua classe pela distância ao centroide do cluster.

  —**Arquitetura 2: Segmentação por Seção + Classificação KMeans**: Semelhante à Arquitetura 1, esta arquitetura também utiliza o KMeans para agrupar textos com base em suas embeddings, a fim de identificar os clusters com maior número de textos de interesse. A diferença é que, nesta abordagem, os textos são segmentados por seção em vez de por sentença. Da mesma forma, a classe de uma nova instância pode ser determinada pela distância ao centroide do cluster.

  —**Arquitetura 3: Segmentação por Sentença + Classificação com BERTimbau**: Nesta arquitetura, os textos são segmentados por sentença, e as embeddings geradas são utilizadas como entrada para o modelo BERTimbau, que realiza a classificação. O BERTimbau é responsável por identificar a classe das novas instâncias com base nas embeddings extraídas.

  —**Arquitetura 4: Segmentação por Seção + Classificação com BERTimbau**: Semelhante à Arquitetura 3, as embeddings geradas são utilizadas como entrada para o modelo BERTimbau, que realiza a classificação. No entanto, nesta abordagem, os textos são segmentados por seção em vez de por sentença.

—**Passo 6: Correlação dos Segmentos com Perfis de Interesse**. Nesta etapa, os segmentos de texto são avaliados quanto à sua correspondência com os perfis de interesse. Nesta análise, verificamos se os *clusters* formados pelo *KMeans* ou as categorias atribuídas pelo classificador estão alinhados com os segmentos de interesse (i.e., segmentos de textos onde há informação de produtos, quantidade e valores).

—**Passo 7: LLM realiza sumarização extrativa**. Nesta etapa, utilizamos o modelo de linguagem GPT-3.5 Turbo para efetuar a sumarização de textos. A escolha do GPT-3.5 Turbo para esta tarefa deve-se à sua capacidade de gerar respostas coesas e contextualmente relevantes a partir de uma variedade de entradas textuais. A resposta esperada deve ser uma lista de dicionários, onde cada dicionário representa um produto, contendo chaves para "produto", "quantidade" e "valor". Cada chave deve detalhar o nome do produto, a quantidade solicitada ou a falta dela, e o valor ou a ausência de informação sobre o valor.

5. RESULTADOS

A Tabela 1 apresenta os resultados dos experimentos de segmentação por sentença e por seção, usando os algoritmos *Kmeans* e *BERTimbau*. Reportamos os resultados em relação à média (AVG) e desvio padrão (SD) para cada uma das métricas: precisão, recall e F1-score. Os resultados mostram que o *BERTimbau* superou o *Kmeans* em todas as métricas. Para a segmentação por sentença, o *BERTimbau* alcançou uma precisão média de 0,69 e um *recall* de 0,71, enquanto o *Kmeans* teve uma precisão de 0,47 e um *recall* de 0,62. Na segmentação por seção, o *BERTimbau* atingiu uma precisão de 0,71 e um *recall* de 0,77, em contraste com a precisão de 0,49 e *recall* de 0,64 do *Kmeans*. O *BERTimbau* também obteve *F1 score* superiores, com 0,69 e 0,72 para segmentação por sentença e seção comparados a 0,50 e 0,52 do *Kmeans*. Os valores de desvio padrão para o *Bertimbau* e *Kmeans* são similares, sugerindo que ambas as técnicas estão sujeitas a um nível semelhante de flutuação em seu desempenho.

	Segmentação por sentença		Segmentação por seção	
	Kmeans	Bertimbau	Kmeans	Bertimbau
AVG precisão	0,47	**0,69**	0,49	**0,71**
AVG recall	0,62	**0,71**	0,64	**0,77**
AVG F1_score	0,50	**0,69**	0,52	**0,72**
SD precisão	0,44	0,39	0,43	0,37
SD recall	0,47	0,398	0,44	0,35
SD F1_score	0,45	0,39	0,42	0,35
AVG Acerto valor	0,35	**0,42**	0,42	**0,45**
AVG Acerto quantidade	0,40	**0,42**	0,38	**0,48**
SD Acerto valor	0,41	0,40	0,40	0,46
SD Acerto quantidade	0,46	0,46	0,44	0,41

Table I. Resultados dos experimentos para a segmentação por sentença e por seção.

Em relação aos acertos, tanto em valor quanto em quantidade, o método *Bertimbau* também mostrou resultados superiores. Na Tabela 1 consideramos como "acerto de valor" quantas vezes o valor dos produtos foi extraído de forma correta e como "acerto de quantidade" quantas vezes a quantidade dos produtos foi extraída de forma correta. Na segmentação por seção, por exemplo, *Bertimbau* alcançou uma média de 0,45 em acerto de valor e 0,48 em acerto de quantidade, comparado a 0,42 e 0,38 do *Kmeans*, respectivamente.

6. CONCLUSÃO E TRABALHOS FUTUROS

Este estudo destacou a relevância e a eficácia de integrar técnicas avançadas de processamento de linguagem natural e aprendizado de máquina para a segmentação e sumarização automatizada de textos em editais de licitação. Com a aplicação de modelos avançados, como o GPT-3.5, conseguimos não só extrair informações críticas de forma eficiente mas também facilitar a tomada de decisões mais informadas no âmbito das licitações públicas. Este processo não apenas potencializa a transparência mas também aumenta a eficiência operacional, reduzindo o tempo e o esforço manual necessários na análise de documentos extensos. No entanto, enfrentamos desafios significativos, especialmente relacionados às "alucinações" do modelo, onde informações plausíveis, porém incorretas, são geradas. Esta limitação exige uma supervisão rigorosa e a implementação de etapas de validação para assegurar a integridade e a precisão das informações extraídas. Para melhorar a precisão e a aplicabilidade dos sumários produzidos, é essencial integrar a revisão humana com ajustes técnicos avançados, além de empregar validações automáticas que assegurem a fidelidade ao texto original. Para futuras pesquisas, propõe-se explorar novas técnicas de segmentação como por exemplo a segmentação a nível de páginas, além do desenvolvimento de modelos ainda mais robustos que possam lidar eficazmente com a complexidade dos textos de licitações.

REFERENCES

Ahmed, M., Seraj, R., and Islam, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* 9 (8): 1295, 2020.

ANDRADE, S. and BAPTISTA, C. d. S. Uso de processamento de linguagem natural e aprendizagem de máquina para a extração de informação em editais de licitações não-estruturados. In *Universidade Federal de Campina Grande*. UFCG, 2022.

Awasthi, I., Gupta, K., Bhogal, P. S., Anand, S. S., and Soni, P. K. Natural language processing (nlp) based text summarization - a survey. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. pp. 1310–1317, 2021.

Cho, S., Song, K., Wang, X., Liu, F., and Yu, D. Toward unifying text segmentation and long document summarization. *arXiv preprint arXiv:2210.16422*, 2022.

da República, P. Lei de licitações e contratos administrativos, 2021.

da Silva, F., Guimarães, G. M. C., Marcacini, R. M., Queiroz, A. L., Borges, V. R. P., Faleiros, T. d. P., and Garcia, L. P. F. Named entity recognition approaches applied to legal document segmentation. In *Anais do X Symposium on Knowledge Discovery, Mining and Learning*. SBC, pp. 210–217, 2022.

da Silva, F., Guimarães, G., Marcacini, R., Queiroz, A., Borges, V. R. P., Faleiros, T., and Garcia, L. Named entity recognition approaches applied to legal document segmentation. In *Anais do X Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 210–217, 2022.

da União, C.-G. Portal da transparência, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

dos Santos Chaves, E. Aspectos importantes da fase interna da licitação: uma análise sobre o conjunto de elementos necessários e suficientes para a caracterização do objeto do processo licitatório. *Revista Controle: Doutrinas e artigos* 13 (1): 149–170, 2015.

Glavaš, G., Nanni, F., and Ponzetto, S. P. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, C. Gardent, R. Bernardi, and I. Titov (Eds.). Association for Computational Linguistics, Berlin, Germany, pp. 125–130, 2016.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Hearst, M. A. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.* 23 (1): 33–64, mar, 1997.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing*. Pearson, 2023. In preparation. Draft chapters available at: `https://web.stanford.edu/~jurafsky/slp3/`.

LaValley, M. P. Logistic regression. *Circulation* 117 (18): 2395–2399, 2008.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

McCulloch, W. S. (1943) warren s. mcculloch and walter pitts a logical calculus of the ideas immanent in nervous activity bulletin of mathematical biophysics 5: 115-133. *Bulletin of mathematical biophysics* vol. 5, pp. 115–133, 1943.

OpenAI. Gpt-4 technical report, 2023. Accessed: 2024-09-27.

Russell, S. J. and Norvig, P. *Artificial intelligence: a modern approach*. Pearson, 2016.

Schröer, C., Kruse, F., and Gómez, J. M. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science* vol. 181, pp. 526–534, 2021.

Somvanshi, M., Chavan, P., Tambade, S., and Shinde, S. A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing communication control and automation (ICCUBEA)*. IEEE, pp. 1–7, 2016.

Souza, F., Nogueira, R., and Lotufo, R. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, R. Cerri and R. C. Prati (Eds.). Springer International Publishing, Cham, pp. 403–417, 2020.

Su, X., Yan, X., and Tsai, C.-L. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics* 4 (3): 275–294, 2012.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning* vol. 8, pp. 279–292, 1992.

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., et al. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, 2024.