MODELO DE LINGUAGEM PARA GERAÇÃO AUTOMÁTICA DE MINUTAS

Maria Eduarda Pinheiro Carreiro[1]; Rafael da Costa Fonsêca[2]; Anne Cristie Pinheiro Gaudencio[3]; Oberdan Rocha Pinheiro[4]
1 Bolsista do SENAI CIMATEC; Projeto de desenvolvimento de pesquisa e inovação; maria.carreiro@fbter.org.br.
2 Especialista I no SENAI CIMATEC; Salvador-BA; rafael.f@fieb.org.br.
3 Consultor I no SENAI CIMATEC; Salvador-BA; anne.moreira@fieb.org.br.
4 Especialista III no SENAI CIMATEC; Salvador-BA; oberdan.pinheiro@fieb.org.br.

RESUMO

O projeto propõe a criação de um Sistema Inteligente baseado em Aprendizado de Máquina, utilizando o modelo LLAMA 13b, treinado com dados da Procuradoria Geral do Estado da Bahia (PGE-BA). Com 13 bilhões de parâmetros, o LLAMA é especificamente adaptado para automatizar classificações e gerar minutas em ações judiciais nos juizados especiais. O processo de fine tuning, utilizando dados específicos da PGE-BA, aprimora a compreensão do modelo em relação às nuances jurídicas, fortalecendo também o vocabulário e o estilo de escrita jurídica em língua portuguesa. A implementação visa otimizar o trâmite processual, agilizando e melhorando a eficiência do sistema jurídico da PGE-BA. A ênfase na geração automática de minutas destaca-se como uma ferramenta valiosa para os profissionais, economizando tempo na produção de documentos legais. No desenvolvimento do projeto, técnicas como a sumarização (Bert Summarizer) e o GPT-2 (Generative Pre-trained Transformer) foram utilizadas, porém o LLAMA revelou-se mais adequado.

PALAVRAS-CHAVE: Geração de texto jurídico, LLM's, Inteligência Artificial, Aprendizado de Máquina.

1. INTRODUÇÃO

A inteligência artificial (IA) tem desempenhado um papel cada vez mais proeminente em diversas áreas, e o campo da geração automática de texto não é exceção. Este projeto tem como objetivo desenvolver e apoiar a implementação de um Sistema Inteligente baseado em Aprendizado de Máquina para automatizar classificações e fornece sugestões durante a tramitação de ações judiciais de demandas em massa nos juizados especiais da Procuradoria Geral do Estado (PGE-BA). O sistema consistirá em três modelos de IA: 1) Modelo de Clusterização de temas processuais; 2) Modelo de Classificação do assunto do processo; e 3) Geração de Minutas de peças processuais, sendo este último o foco deste relatório.

Este modelo tem como objetivo principal gerar minutas para diversos tipos de documentos jurídicos, abrangendo uma ampla gama de temas. As dificuldades encontradas durante seu desenvolvimento incluem a necessidade de manter o contexto dos documentos utilizados como inputs e fundamentar o conteúdo legalmente com base no arcabouço de leis utilizadas no treinamento do modelo. Adicionalmente, busca-se manter uma excelente qualidade de escrita, empregando corretamente o idioma português e o vocabulário jurídico específico.

2. METODOLOGIA

Após a exploração das diversas abordagens tecnológicas, vale destacar que os Large Language Models (LLMs) representam uma categoria proeminente de modelos de linguagem utilizados neste projeto. Esses LLMs, como GPT[1,] Palm[2] e Llama[3] são sistemas de inteligência artificial capazes de compreender e gerar texto em larga escala, com base em um extenso treinamento em dados linguísticos. Sua capacidade de entender e produzir linguagem natural com nuances semelhantes às humanas os torna fundamentais em uma variedade de aplicações, incluindo a geração de minutas jurídicas. Os LLMs têm sido objeto de intensa pesquisa e desenvolvimento, evoluindo continuamente para oferecer desempenho aprimorado e atender às demandas de diversas áreas, como o direito.

Além da exploração dos Large Language Models (LLMs), foram conduzidos testes adicionais utilizando modelos como o T5[4] e redes GANs.[5] Esses modelos representam diferentes abordagens tecnológicas que ofereceram insights sobre as capacidades e limitações de diferentes estratégias de geração de texto. O T5, por exemplo, é capaz de realizar tarefas de linguagem natural com uma abordagem de "text-to-text", enquanto as redes GANs são reconhecidas por sua capacidade de gerar dados sintéticos realistas por meio de um processo de aprendizado adversarial. Essa diversidade de testes permitiu uma compreensão mais completa das opções disponíveis, sendo fundamental para a seleção do modelo mais adequado às necessidades específicas do projeto.

Após uma extensa avaliação, o modelo LLAMA 2 se destacou como a escolha mais promissora, especialmente sua versão com 13 bilhões de parâmetros do tipo chat, oferecendo um equilíbrio notável entre desempenho e custo computacional.

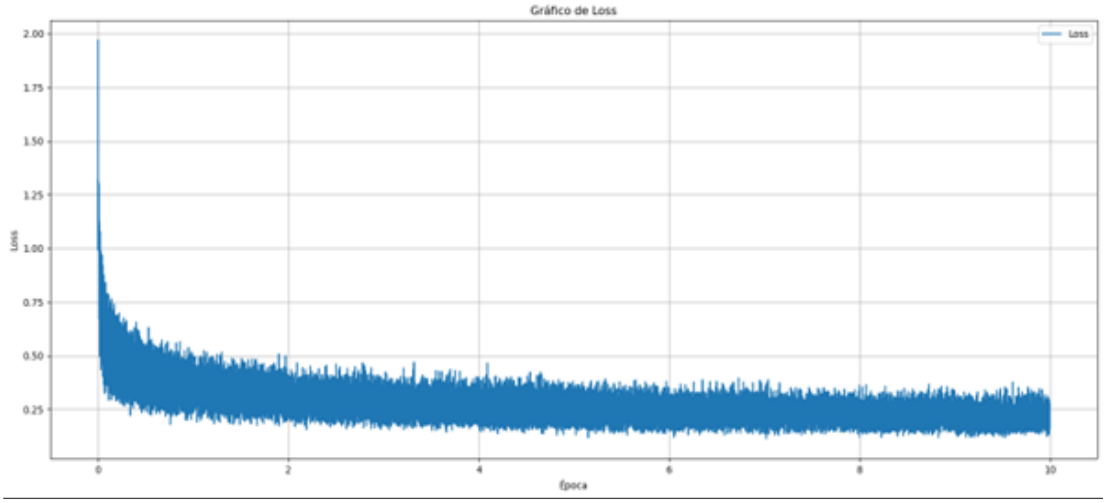
No processo de avaliação para a escolha do modelo mais adequado do LLAMA 2, foram considerados diversos fatores, incluindo desempenho e custo computacional. Em relação ao modelo 7b, com seus 7 bilhões de parâmetros, embora ofereça uma vantagem em termos de custo computacional menor em comparação com o modelo de 13 bilhões de parâmetros, sua performance é consideravelmente inferior. Optou-se, portanto, por não utilizar o modelo 7b devido à sua capacidade limitada de geração de texto, priorizando a qualidade e eficácia na produção das minutas jurídicas.

Por outro lado, o modelo 70b, com 70 bilhões de parâmetros, representa uma opção mais robusta e poderosa em termos de capacidade de geração de texto. No entanto, essa vantagem é acompanhada por um custo computacional significativamente mais elevado em comparação com o modelo de 13 bilhões de parâmetros. Considerando a necessidade de equilibrar eficiência e custo, optou-se por não utilizar o modelo 70b. Embora oferecesse uma performance superior, os recursos computacionais adicionais necessários para sua implementação seriam excessivos em relação aos benefícios adicionais proporcionados, tornando o modelo de 13 bilhões de parâmetros a escolha mais equilibrada para atender às necessidades do projeto. A precisão e eficiência demonstradas pelo Llama 2, aliadas à sua capacidade de adaptação às particularidades do contexto jurídico da PGE-BA, corroboram sua seleção como a solução ideal para a geração de minutas jurídicas nesse ambiente específico.

No entanto, dada a especificidade do contexto jurídico da Procuradoria Geral do Estado da Bahia (PGE-BA), optou-se por implementar um ajuste fino no modelo selecionado, o Llama 2 com 13 bilhões de parâmetros. Esse ajuste fino envolveu a utilização de um conjunto de dados próprio, composto por documentos jurídicos emitidos pela PGE-BA e a legislação baiana relevante. O processo de fine tuning exigiu aproximadamente 15 dias de treinamento, utilizando 4 GPUs, com dataset próprio de 201 mil linhas contendo os documentos jurídicos emitidos pela PGE-BA e a legislação baiana. O processo resultou em um modelo adaptado às nuances e particularidades locais. Essa etapa foi crucial para garantir que o modelo Llama 2 atendesse de forma precisa e eficiente às demandas específicas do ambiente jurídico da PGE-BA, garantindo, assim, uma geração de minutas jurídicas de alta qualidade e relevância para a instituição.

3. RESULTADOS E DISCUSSÃO

O treinamento de ajuste fino mencionado anteriormente, utilizando dados jurídicos, exibiu o comportamento demonstrado pelo gráfico abaixo, que representa a evolução da função de perda durante o processo de treinamento. Conforme observado, o treinamento foi composto por 10 épocas e concluiu com o valor de perda (loss) inferior a 0.4.



O modelo de Geração de Minutas Processuais desenvolvido apresentou bons resultados, sendo capaz de gerar minuta de 4 tipos diferentes de documentos jurídicos de diversos temas, com contexto e qualidade de escrita adequados. A avaliação desta é realizada qualitativamente, através da comparação com os documentos jurídicos já existentes da Procuradoria Geral do Estado da Bahia (PGE-BA). Esse processo envolveu análises minuciosas e contínuas melhorias até alcançar o estado atual do modelo, garantindo que as minutas geradas atendessem aos padrões de qualidade e contexto exigidos pela PGE-BA.
Outro aspecto essencial para o bom desempenho do modelo foi a aplicação de Engenharia de Prompt, definida como a arte e a ciência de criar entradas (prompts) para obter os resultados desejados de modelos de IA maximizando assim a eficácia desses modelos.[6]

Isso inclui a formulação de instruções específicas sobre o que se espera do texto gerado, a definição do persona, como um advogado experiente, e exemplos do texto que o modelo deve ser capaz de produzir.

4. CONSIDERAÇÕES FINAIS

Os textos gerados até o momento demonstram uma qualidade notável, refletindo não apenas a precisão técnica do modelo, mas também sua capacidade de contextualização e adequação ao propósito específico, como evidenciado pela síntese processual e conclusão fornecidas. A habilidade do modelo em produzir textos coerentes e relevantes para um contexto jurídico, como o requerido pela PGE-BA, é notável. No entanto, ainda há espaço para implementações adicionais visando aprimorar ainda mais sua utilidade e versatilidade. Por exemplo, a inclusão de tópicos específicos, que pode expandir sua aplicabilidade em diferentes áreas do direito. Além disso, a melhoria contínua da capacidade de compreensão e interpretação do modelo, especialmente em relação a nuances legais e terminologia especializada, contribuiria para um desempenho ainda mais excepcional. Com essas melhorias, o modelo poderá oferecer resultados ainda mais precisos e abrangentes, atendendo às demandas de um espectro mais amplo de usuários e cenários.

Agradecimentos

Obrigada a todos pelo trabalho em equipe excepcional, um agradecimento especial ao nosso orientador Oberdan Rocha Pinheiro pelo seu apoio fundamental.

5. REFERÊNCIAS

1 SOLAIMAN, Irene et al. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203, 2019.

2 CHOWDHERY, Aakanksha et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, v. 24, n. 240, p. 1-113, 2023.

3 TOUVRON, Hugo et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

4 RAFFEL, Colin et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, v. 21, n. 1, p. 5485-5551, 2020.

5 CRESWELL, Antonia et al. Generative adversarial networks: An overview. IEEE signal processing magazine, v. 35, n. 1, p. 53-65, 2018.

6 MITTAL, Aayush. Guia De Engenharia Imediata Da Openai: Dominando O Chatgpt Para Aplicativos Avançados. Unite, 2023. Disponível em: < https://www.unite.ai/pt/openais-prompt-guia-de-engenharia-dominando-chatgpt-para-aplica%C3%A7%C3%B5es-avan%C3%A7adas/>.