Reconhecimento de Entidades Nomeadas (NER): Avaliação de modelos Transformers no domínio jurídico

[1] Paloma Corrêa Alves
Centro de Informática - UFPE, Recife, Brazil
pca2@cin.ufpe.br

***Abstract*—O presente estudo avalia o desempenho de quatro modelos transformers pré-treinados (BERT-base, BERT-large, RoBERTa-base e RoBERTa-large) na tarefa de Reconhecimento de Entidades Nomeadas (NER) em textos jurídicos do corpus LeNER-Br. O dataset foi particionado para treino, validação e teste, e cada modelo foi submetido a fine-tuning com otimizador, treinamento por 10 épocas e early stopping baseado na métrica F1 de validação. Os melhores checkpoints de cada experimento foram versionados no Hugging Face Hub e consumidos em notebooks de token classification, que realizaram a tokenização, o alinhamento preciso de tokens com rótulos e a avaliação com as bibliotecas Evaluate e Seqeval. Os resultados mostram que o BERT-large alcançou o melhor F1 global, seguido pelo BERT-base, enquanto RoBERTa-base e RoBERTa-large apresentaram menores percentuais de F1, evidenciando um trade-off entre capacidade do modelo e tamanho do dataset. Para demonstrar a aplicabilidade prática, foi desenvolvida uma interface web em Gradio que destaca automaticamente as entidades reconhecidas no texto jurídico submetido pelo usuário.**

Palavras-chave: Processamento da Linguagem Natural, Reconhecimento de Entidades nomeadas, Transformers.

I. Introdução

O crescimento acelerado dos dados textuais no meio jurídico impõe desafios significativos para a extração automatizada de informações essenciais. Nesse contexto, as técnicas de Processamento de Linguagem Natural (PLN) tornaram-se ferramentas indispensáveis para estruturar e analisar tais informações. Em particular, o Reconhecimento de Entidades Nomeadas (NER) se destaca por sua capacidade de identificar e categorizar elementos específicos, como referências a leis, órgãos judiciais e partes envolvidas, presentes nos documentos legais.

O reconhecimento de entidades nomeadas, conhecido pela sigla NER, é um ramo do processamento de linguagem natural que se dedica à extração e classificação de entidades presentes em textos não estruturados. Essas entidades são organizadas em categorias pré-definidas que representam conceitos ou objetos específicos de um determinado domínio, tais como pessoas, lugares, organizações, normas e autoridades (MÔRO, 2018). Dessa forma, o NER se mostra uma técnica valiosa para extrair e analisar informações relevantes em documentos especializados, como os jurídicos.

Em 2018, a Google introduziu uma inovação que transformaria o cenário do PLN com o desenvolvimento do BERT (Bidirectional Encoder Representations from Transformers). Essa abordagem, baseada em pré-treinamento seguido de fine-tuning, revolucionou a área ao proporcionar uma representação do contexto de forma bidirecional. Conforme destacado por Devlin et al. (2019), essa característica permite ao modelo captar nuances contextuais tanto à esquerda quanto à direita de cada token, superando as limitações inerentes aos modelos unidirecionais que não conseguem integrar completamente o contexto em nível de token.

Reconhecendo a necessidade de adaptações para o português, especialmente no ambiente jurídico, pesquisadores e empresas especializadas passaram a desenvolver variantes do BERT focadas e adaptadas ao português, que rapidamente se destacaram ao atingir resultados de ponta na tarefa de NER (SOUZA et al., 2019). Paralelamente, arquiteturas como BERT-base-cased e RoBERTa têm mostrado que, ao fornecer representações contextuais refinadas, é possível ajustar esses modelos para tarefas específicas de maneira bastante eficaz, por meio de técnicas de fine-tuning.

A. Contextualização do BERT

O BERT (Bidirectional Encoder Representations from Transformers) revolucionou o campo do Processamento de Linguagem Natural (PLN) ao introduzir uma abordagem pré-treinada baseada em transformadores capaz de capturar contextos de forma bidirecional. O BERT rapidamente se estabeleceu como referência para diversas tarefas de PLN, tais como classificação de texto, resposta a perguntas, tradução e reconhecimento de entidades nomeadas (NER). Este modelo, fundamentado em redes neurais artificiais, destaca-se pelo desempenho excepcional em uma ampla variedade de aplicações de processamento de textos.

A arquitetura "base" do BERT, conhecida como BERT-base, é composta por 12 camadas de transformadores, 768 dimensões de embeddings e aproximadamente 110 milhões de parâmetros, proporcionando um equilíbrio entre capacidade de modelagem e exigências computacionais para uma ampla

gama de aplicações. O BERT é uma arquitetura de rede neural baseada em transformer que usa apenas a parte do encoder da arquitetura para gerar um modelo da linguagem. Esse modelo é aprendido durante o pré-treinamento com grande quantidade de textos no idioma desejado. Em comparação com a arquitetura original do transformer, o modelo BERT-base possui 12 camadas no codificador, enquanto o modelo BERT-large possui 24 camadas (KENTON; TOUTANOVA, 2019).

Uma rede BERT é pré-treinada em duas tarefas, Masked-Language Modeling (MLM) e Next Sentence Prediction (NSP). No MLM, uma porcentagem dos tokens são escolhidos aleatoriamente antes das sequências de texto serem usadas como entrada da rede no pré-treino (DEVLIN et al., 2018). Depois do pré-treino, o ajuste fino é realizado levando em conta a tarefa final para que o modelo será utilizado, como NER no caso deste trabalho. Comparado ao pré-treinamento, que é bastante custoso, o ajuste fino é pouco dispendioso e pode ser realizado em poucas horas numa TPU ou GPU (DEVLIN et al., 2018).

A arquitetura BERT-base possui duas variantes principais que se diferenciam, sobretudo, pelo tratamento de maiúsculas e minúsculas, dando origem aos modelos "uncased" e "cased". O modelo "BERT-base-uncased" foi projetado com a premissa de que a normalização do texto para minúsculas pode reduzir a complexidade do vocabulário, facilitando o treinamento e, muitas vezes, melhorando a generalização em tarefas onde a distinção entre maiúsculas e minúsculas não é fundamental.

Esse pré-processamento pode ser particularmente útil em domínios em que a capitalização é inconsistente, como em textos extraídos de redes sociais ou documentos que não seguem rigorosamente as normas gramaticais. Por outro lado, o "BERT-base-cased" preserva as informações originais do texto, permitindo que o modelo capture nuances semânticas que dependem da capitalização – uma característica essencial em textos formais, literários ou em domínios onde nomes próprios e siglas têm grande relevância. Essa distinção, embora sutil, pode impactar significativamente o desempenho do modelo em tarefas que exigem uma compreensão mais refinada do contexto linguístico.

II. Solução Proposta

O Reconhecimento de Entidades Nomeadas (NER) no domínio jurídico é um problema desafiador de Processamento de Linguagem Natural, cujo objetivo é identificar automaticamente em textos legais as menções a entidades relevantes, como nomes de pessoas, organizações, lugares, expressões temporais, legislações e casos jurídicos (jurisprudências). A motivação para este projeto surge da necessidade de simplificar a extração de informações em documentos extensos e complexos, auxiliando profissionais do direito a localizar rapidamente referências importantes em leis, decisões judiciais e outros documentos legais. A solução proposta busca desenvolver um sistema de NER especializado para o contexto jurídico brasileiro, capaz de etiquetar automaticamente essas entidades em textos em português, facilitando a análise jurídica.

Para abordar esse problema, foram utilizados modelos do tipo Transformer pré-treinados que representam o estado da arte em NER. Em particular, foram avaliados quatro modelos: BERT-base, BERT-large, RoBERTa-base e RoBERTa-large. Os modelos BERT (Bidirectional Encoder Representations from Transformers) foram escolhidos por seu sucesso em tarefas de NER gerais; neste trabalho foram empregados versões já treinadas em português (BERTimbau). Os modelos RoBERTa utilizados são variações otimizadas do BERT nas versões base e large. O RoBERTa-base conta com 12 camadas Transformer e foi pré-treinado em grandes corpos, usando Masked Language Modeling, o que oferece um bom equilíbrio entre desempenho e eficiência computacional em tarefas de NLP. Já o RoBERTa-large amplia essa configuração para 24 camadas e passa por mais etapas de pré-treinamento nos mesmos dados, resultando em ganhos expressivos em benchmarks complexos, embora exija maior memória e tempo de inferência. Ao serem ajustados no dataset LeNER-Br, originaram-se os checkpoints `/ner-bert-lenerbr` (base e large) e `/ner-roberta-lenerbr` (base e large), cada um capaz de rotular automaticamente os tokens do texto de entrada segundo as classes de entidade definidas.

Além do treinamento dos modelos de NER, para demonstrar a aplicabilidade prática, foi desenvolvida uma interface web interativa (web app) usando a biblioteca Gradio (Figura 1). Essa aplicação permite permite aos usuários inserir textos jurídicos e receber como saída o mesmo texto com as entidades identificadas (como pessoas, legislações, jurisprudências, etc) destacadas em cores ou ícones, indicando a qual categoria cada menção pertence. A Figura 1 ilustra a interface web app com um exemplo de entrada e as entidades reconhecidas destacadas no texto. Esse componente de software demonstra a usabilidade da solução, permitindo que usuários finais testem o modelo e visualizem os resultados de forma intuitiva, sem necessidade de conhecimento de programação. Em resumo, a solução proposta abrange tanto a investigação experimental dos modelos Transformers no contexto de NER jurídico quanto a implementação de uma ferramenta acessível que aplica o melhor modelo obtido em um cenário real.

III. Método e Processos de Análise

A. Dataset de Treinamento

Foi utilizado o conjunto LeNER-Br (Legal Named Entity Recognition – Brasil), composto por 70 documentos legais (decisões judiciais e legislações), em português, provenientes de diversos tribunais brasileiros, além de documentos legislativos, manualmente anotados com rótulos que representam seis categorias de entidades relevantes ao contexto jurídico brasileiro (ARAUJO et al., 2018). Este corpus reúne textos provenientes de tribunais de diversas regiões e de diferentes normativos legislativos, garantindo ampla cobertura do jargão e das estruturas textuais do domínio. As seis classes de entidades definidas no LeNER-Br são:

- Organização: nomes de instituições e órgãos (por ex., tribunais, ministérios);

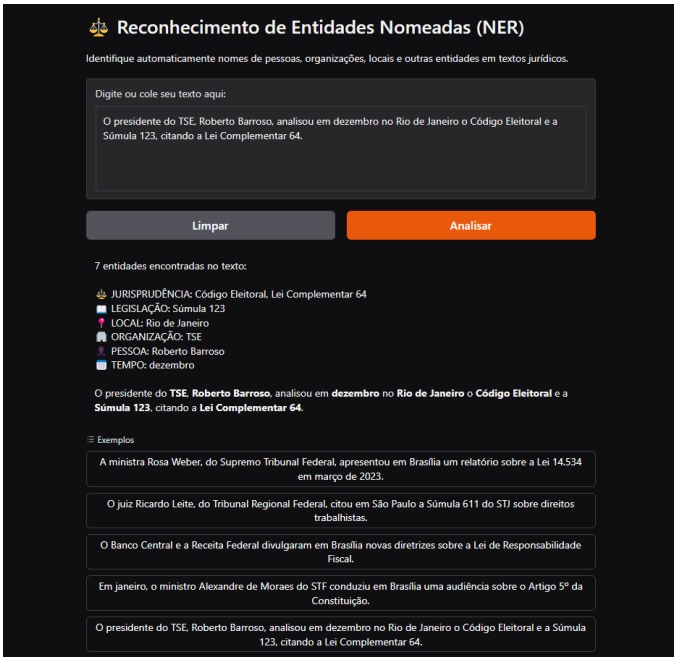

Fig. 1. Interface da aplicação web desenvolvida (Gradio), exibindo um exemplo de texto jurídico de entrada e as entidades nomeadas reconhecidas pelo modelo, destacadas com cores e ícones.

- Pessoa: nomes próprios de indivíduos (por ex., magistrados, partes envolvidas);
- Tempo: expressões temporais (por ex., datas, períodos);
- Local: referências geográficas (por ex., cidades, estados);
- Legislação: menções a leis e normas (por ex., artigos, códigos);
- Jurisprudência: citações de decisões judiciais (por ex., acórdãos, súmulas).

Os dados foram estruturados para facilitar a extração automática de informações e a adaptação de modelos de NLP ao domínio jurídico, permitindo comparações diretas com estudos prévios e a avaliação consistente dos novos modelos.

B. Estratégia de treinamento e classificação

O estudo foi realizado em duas etapas. A Etapa 01 foi realizada nos notebooks de treinamento, nos quais o modelo pré-treinado (BERT-base, BERT-large, RoBERTa-base ou RoBERTa-large) foi fine-tuned diretamente sobre o conjunto de dados LeNER-Br. O ambiente foi preparado com a instalação das bibliotecas necessárias (datasets, transformers, evaluate, accelerate, huggingface_hub e seqeval) e o carregamento do corpus LeNER-Br, dividido em treino, validação e teste. Para cada arquitetura, instanciou-se o tokenizer correspondente — assegurando, no caso de RoBERTa, o parâmetro add_prefix_space=True — e o modelo pré-treinado para token classification. Em seguida, uma função de tokenização alinhou tokens e rótulos de entidade, aplicando dataset.map com batched=True e usando um DataCollatorForTokenClassification.

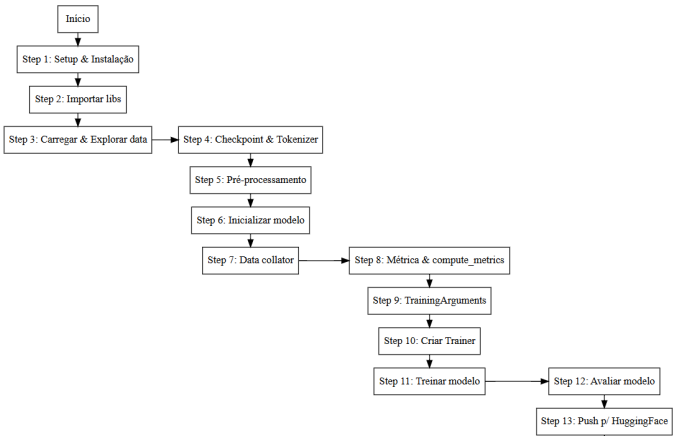

Fig. 2. Etapa 01- Fluxo de atividades do pipeline de treinamento do modelo no dataset LeNER

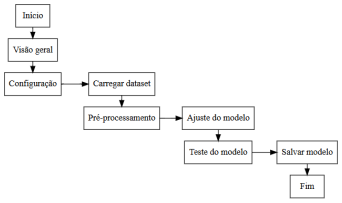

Fig. 3. Etapa 02 - Fluxo de atividades do pipeline de reconhecimento de entidades (NER)

Com os dados prontos, definiu-se um conjunto de TrainingArguments para 10 épocas, batch size de 16, learning rate entre 2e-5 e 5e-5 e early stopping baseado em F1 na validação. O fine-tuning foi conduzido via Trainer.train(), seguido de Trainer.evaluate() para coleta das métricas de precisão, recall e F1-score. Ao final do processo de fine-tuning, o melhor checkpoint obtido foi empurrado automaticamente para o Hugging Face Hub, criando um modelo versionado e acessível por meio de um identificador único.

Durante os experimentos, observou-se que os modelos de maior capacidade — especialmente o BERT-large e o RoBERTa-large — atingiam o pico de F1-score por volta da 5ª ou 6ª época. A partir desse ponto, o desempenho no conjunto de validação começava a cair, indicando o início de sobreajuste. A aplicação do *early stopping* foi, portanto, crucial: ao interromper o treinamento automaticamente no momento de melhor desempenho, evitou-se uma perda de até 1–2 pontos percentuais no F1 final. Esse comportamento reforça a sensibilidade desses modelos ao tamanho relativamente reduzido do corpus LeNER-Br, evidenciando que arquiteturas com maior número de parâmetros exigem estratégias rigorosas de regularização para preservar a capacidade de generalização.

Em seguida, na Etapa 02, foram utilizados os checkpoint nos notebooks de token classification. Ao iniciar cada um deles, o modelo saiu do Hub já ajustado, dispensando a necessidade de retreinar nada localmente. Esses notebooks dedicaram-se a demonstrar o pipeline de NER completo: carregamento do checkpoint, tokenização com o tokenizer apropriado (incluindo

TABLE I
RESULTADOS DE DESEMPENHO DOS MODELOS FINE-TUNED NO CONJUNTO LENER-BR – ETAPA 01

Modelo	Precisão (%)	Recall (%)	F1-score (%)	Acurácia (%)
BERT-base	85,9	89,9	87,9	96,9
BERT-large	89,7	91,0	90,3	98,0
RoBERTa-base	83,8	87,0	85,4	97,2
RoBERTa-large	83,1	89,9	86,4	97,2

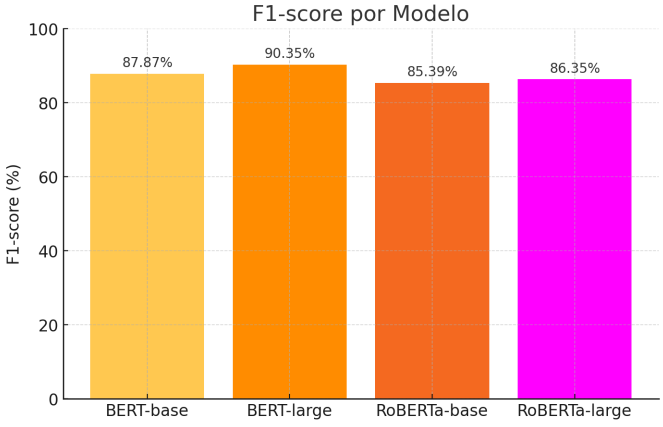

Fig. 4. Comparação do F1-score para os quatro modelos Transformers avaliados (BERT-base, BERT-large, RoBERTa-base e RoBERTa-large). Cada barra representa um modelo, e a altura da barra corresponde ao F1 obtido pelo modelo na Etapa 01.

o add_prefix_space=True para RoBERTa), alinhamento de tokens com rótulos de entidade e avaliação do desempenho. O add_prefix_space= True muda a tokenização e, consequentemente, o alinhamento entre os tokens gerados e o vocabulário em que o modelo foi pré-treinado. O WordPiece do BERT foi mantido em sua configuração original de tokenização (sem prefixar espaço), enquanto o BPE do RoBERTa precisou desse tweak para respeitar onde começam as palavras. Isso explica por que, ao tratar todos os modelos da mesma forma, houve degradação do desempenho dos BERTs.

Como o treinamento já havia sido realizado, a ênfase recaiu sobre a aplicação didática das funções de pré-processamento e sobre a geração de métricas precisas de precisão, recall e F1-score.

Dessa forma, o fluxo de trabalho se desdobrou em duas etapas bem definidas: primeiro, o ajuste fino e a publicação do modelo; depois, o consumo desse modelo em um pipeline completo de token classification. Essa separação garantiu maior modularidade — o mesmo checkpoint pôde ser testado em diferentes cenários ou comparado rapidamente com outras versões — e facilitou tanto a reprodutibilidade dos resultados quanto a compartimentalização entre fases de treinamento e avaliação.

No caso do BERT-large e RoBERTa-large, foi necessário cuidar do sobreajuste (overfitting), dado que modelos com maior número de parâmetros tendem a memorizar o conjunto de treino devido ao tamanho relativamente pequeno do corpus. De fato, foi observado que o treinamento frequentemente atingia o melhor desempenho antes da última época, disparando o critério de parada antecipada. Isso é consistente com outros experimentos relatados devido ao pequeno tamanho do dataset de fine-tuning, o modelo sofreu overfitting antes do final do treinamento. Para mitigar esse efeito, foram aplicadas técnicas como dropout e verificação de desempenho a cada época, salvando apenas os pesos do modelo com melhor F1 na validação.

C. Infraestrutura e Ferramentas

Os experimentos foram conduzidos no Google Colab, aproveitando seus recursos computacionais robustos e o suporte a GPUs para treinamento de deep learning. Esse ambiente facilitou a implementação das rotinas de pré-processamento e a integração prática de modelos pré-treinados, além de permitir a manipulação eficiente do dataset. Para gerenciar cargas e transformar os dados, foram utilizadas as bibliotecas Transformers e Dataset da Hugging Face, que simplificam o download de checkpoints e a preparação de batches. Na etapa de avaliação, foram adotadas as ferramentas Evaluate e Seqeval, obtendo métricas detalhadas de precisão, recall, F1-score e acurácia.

Complementando o fluxo de trabalho, foi empreagado Accelerate para orquestrar o treinamento distribuído e o Hugging Face Hub para versionamento automático dos modelos, garantindo rastreabilidade e reprodutibilidade. Nos testes com os modelos base (BERT-base e RoBERTa-base), uma GPU Tesla T4 foi suficiente para concluir o treinamento em poucas horas. Já as variantes large, com centenas de milhões de parâmetros, demandaram uma NVIDIA A100 de 40 GB (via Colab Pro) e apresentaram tempos de execução significativamente maiores.

Cada notebook de treinamento registrou as métricas por época e ao final gerou o relatório de avaliação no conjunto de validação. As métricas utilizadas para comparação foram a precisão, recall e F1-score por classe de entidade e no geral. O F1-score, sendo a média harmônica de precisão e recall, é a principal métrica para avaliar o equilíbrio entre falsos positivos e falsos negativos na detecção de entidades.

IV. RESULTADOS

Nesta seção será apresentado quantitativamente o desempenho obtido por cada modelo. A Tabela 2 resume as métricas globais de precisão, recall, F1-score e acurácia alcançadas no conjunto de avaliação para os quatro transformadores fine-tunados. Já a Tabela 3 detalha os F1-scores por categoria de entidade para cada modelo, evidenciando diferenças de desempenho em tipos específicos. Conforme esperado, os modelos Large tendem a superar suas versões Base em desempenho, embora com diferenças distintas em cada arquitetura.

Analisando os resultados globais (Tabela 1), observa-se que o BERT-large obteve a melhor performance geral, com $F_1 \approx 90,0\%$, seguido de perto pelo BERT-base e Roberta-large, ambos com $F_1 \approx 88,6\%$. O modelo RoBERTa-base alcançou um desempenho um pouco inferior: obteve $F_1 \approx 85,3\%$, enquanto o RoBERTa-large teve $F_1 \approx 88,6\%$.

TABLE II
RESULTADOS DE DESEMPENHO DOS MODELOS – ETAPA 02

Modelo	Precisão (%)	Recall (%)	F1-score (%)	Acurácia (%
BERT-base	88,5	88,8	88,6	96,8
BERT-large	88,6	91,5	90,0	97,9
RoBERTa-base	84,6	86,1	85,3	97,3
RoBERTa-large	86,8	90,4	88,6	97,6

TABLE III
F1-SCORE (%) POR TIPO DE ENTIDADE NOMEADA PARA CADA MODELO AVALIADO

Categoria	BERT-base	BERT-large	RoBERTa-base	RoBERTa-large
JURISPRUDÊNCIA	86,1	82,8	90,2	98,1
LEGISLAÇÃO	92,9	95,3	90,9	92,5
LOCAL	65,9	68,3	59,4	65,4
ORGANIZAÇÃO	85,3	88,1	82,4	86,4
PESSOA	96,8	97,5	91,4	91,5
TEMPO	95,6	95,6	93,3	94,8

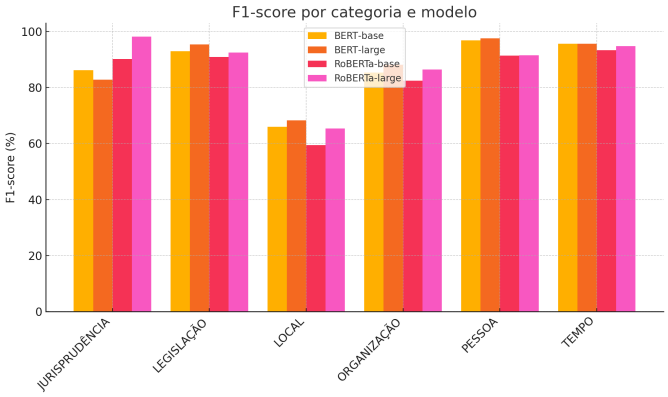

Fig. 5. Comparação do F1-score por categoria de entidade para os quatro modelos Transformers avaliados (BERT-base, BERT-large, RoBERTa-base e RoBERTa-large). Cada grupo de barras representa uma categoria, e a altura da barra corresponde ao F1 obtido pelo modelo nessa classe

Outro aspecto relevante é o trade-off entre recall, nota-se que o BERT-large equilibrou bem ambos (precisão 88,6% vs recall 91,5%), ao passo que o RoBERTa-large apresentou um desequilíbrio maior (precisão 86,8% vs recall 90,4%), indicando maior taxa de falsos positivos nesse modelo. Em aplicações jurídicas, esse comportamento pode ser problemático, pois um excesso de falsos positivos pode levar a alertas indevidos e comprometer a confiabilidade do sistema. Já o BERT-base e RoBERTa-base exibiram valores de precisão e recall mais próximos entre si, sugerindo comportamento mais estável, sugerindo uma detecção mais conservadora e precisa. Essa diferença pode ser atribuída ao fato de o BERT-large ter sido pré-treinado em grandes corpora de português, capturando melhor as nuances do idioma jurídico brasileiro, enquanto o RoBERTa-large, pode ter sofrido com ruídos linguísticos no pré-treinamento.

No detalhamento por classe, ressalta-se:

- *PESSOA*: F1 elevado para todos os modelos, de 91,4% (`RoBERTa-base`) a 97,5% (`BERT-large`), reflexo da alta frequência desses exemplos.

- *TEMPO*: F1 variando de 93,3% (`RoBERTa-base`) a 95,6% (`BERT-base` e `BERT-large`), indicando reconhecimento consistente de padrões numéricos e temporais.

- *LOCAL*: apresentou os menores F1-scores, de 59,4% (`RoBERTa-base`) a 68,3% (`BERT-large`), apontando maior dificuldade na identificação de toponímicos. Apesar do desempenho global elevado, o melhor modelo avaliado (BERT-large) atingiu apenas cerca de 68% de F1 na classe *LOCAL*. Esse resultado reflete a dificuldade particular dessa categoria, que representa menos de 1% dos exemplos no corpus e apresenta alta ambiguidade, com nomes de locais muitas vezes confundidos com nomes de pessoas. Tal limitação evidencia que simplesmente aumentar a capacidade do modelo não é suficiente para lidar com a escassez de exemplos: seria necessário aplicar técnicas de *data augmentation* específicas, como o uso de *gazetteers* de topônimos, para enriquecer o conjunto de dados e melhorar a cobertura dessa entidade.

- *LEGISLAÇÃO* e *ORGANIZAÇÃO*: os modelos BERT mostraram leve superioridade (F1 de 92,9%–95,3% e 85,3%–88,1%, respectivamente) em comparação aos RoBERTa (90,9%–92,5% e 82,4%–86,4%). A comparação entre base e large evidencia os limites impostos pela escassez de dados. Apesar de o modelo large dobrar o número de camadas em relação à versão base, os ganhos em categorias como LEGISLAÇÃO e ORGANIZAÇÃO não foram tão altos. Isso sugere que, sem o acréscimo de mais exemplos específicos durante o fine-tuning, especialmente em domínios especializados como o jurídico, modelos maiores tendem a estagnar. Esse comportamento reforça a ideia de que o aumento da capacidade do modelo, por si só, não garante melhorias significativas se o conjunto de dados não for suficientemente representativo ou diversificado.

- *JURISPRUDÊNCIA*: variação de 82,8% (`BERT-large`) a 98,1% (`RoBERTa-large`), sugerindo que o `RoBERTa-large` se beneficiou mais desta classe em particular.

Essa diferença indica que os modelos RoBERTa tiveram dificuldade maior em generalizar para nomes de lugares, talvez devido à menor representação destes no corpus ou diferenças nos embeddings de língua para toponímia. De fato, no corpus LeNER-Br a frequência de locais é bem menor que a de outras entidades (apenas 0,28% das entidades no teste são locais e muitos erros decorrem de confusão entre nomes de pessoas e locais.

Outra discrepância notável está nas classes LEGISLAÇÃO e JURISPRUDÊNCIA, exclusivas do domínio jurídico. Em LEGISLAÇÃO (referências a leis, constituições etc.), os modelos BERT alcançaram F1 de 92,9% (BERT-base) a 95,3% (BERT-large), enquanto as variantes RoBERTa obtiveram 90,9% (RoBERTa-base) e 92,5% (RoBERTa-large).

Para JURISPRUDÊNCIA (casos judiciais, decisões), o comportamento foi diferente, RoBERTa-large brilhou em JURISPRUDÊNCIA, mas perdeu em LEGISLAÇÃO, enquanto o BERT-large levou vantagem em LEGISLAÇÃO. O RoBERTa-large chegou a 98,1 % em JURISPRUDÊNCIA — um ganho

de 10 pp sobre o BERT-large nessa classe.

O desempenho superior do RoBERTa-large na classe *JURISPRUDÊNCIA* sugere que sua estratégia de mascaramento mais agressiva durante o pré-treinamento favoreceu a captura de padrões linguísticos complexos, como estruturas recorrentes em citações de acórdãos. No entanto, esse ganho não se estendeu à classe *LEGISLAÇÃO*, onde o modelo não obteve o mesmo destaque. Esse resultado pode ser atribuído à sua capacidade ampliada de modelar expressões técnicas e formulações típicas do discurso jurídico, beneficiando especialmente o reconhecimento de referências jurisprudenciais. Entretanto, no caso do BERT esse mesmo poder de modelagem parece ter levado a um leve sobreajuste (overfitting) para exemplos de jurisprudência, pois o BERT-large, embora maior que o BERT-base, obteve desempenho inferior (82,8% contra 86,1%) nessa categoria.

Embora o F1-score global (micro) obtido pelo BERT-large tenha alcançado cerca de 90%, essa métrica tende a ser inflada pelas classes mais frequentes e de fácil reconhecimento, como *PESSOA* e *TEMPO*. Ao se considerar uma média macro, que atribui o mesmo peso a todas as classes independentemente de sua frequência, o desempenho médio geral dos modelos tende a cair para aproximadamente 85%. Essa queda se deve especialmente ao baixo desempenho nas classes *LOCAL* e *JURISPRUDÊNCIA*, que apresentaram maior dificuldade de generalização. Esse aspecto é crucial quando se busca uma cobertura uniforme entre todas as entidades, especialmente em aplicações onde a detecção precisa de categorias menos frequentes é tão importante quanto das mais recorrentes.

Para visualizar comparativamente os desempenhos por categoria de entidade, a Figura 2 apresenta um gráfico de barras com os F1-scores de cada modelo em cada tipo de entidade. Observa-se claramente, por exemplo, a superioridade do BERT-large em quase todas as categorias (barras laranja), bem como a queda acentuada dos quatro modelos em LOCAL. Gráficos como este ajudam a evidenciar os pontos fortes e fracos de cada abordagem em nível de entidade, complementando os valores numéricos das tabelas.

Em resumo, os experimentos mostraram que BERT-large foi o modelo com melhor desempenho geral em NER jurídico, seguido de perto por BERT-base. As variantes RoBERTa não superaram os modelos BERT, e em particular a versão base do RoBERTa enfrentou dificuldades de generalização no conjunto de validação.

V. Discussão

Os resultados obtidos permitem uma análise sobre o uso de Transformers em NER jurídico. Em primeiro lugar, foram constatadas que modelos pré-treinados em português, como o BERTimbau, oferecem uma base mais sólida para essa tarefa – os dois modelos BERT avaliados superaram os modelos RoBERTa correspondentes. Uma possível explicação é que o BERT utilizado foi explicitamente treinado em grandes corpora de português jurídico (ou ao menos português geral), enquanto os modelos RoBERTa podem não captar tão bem as nuances do português jurídico.

Trabalhos prévios já indicavam a importância de especializar modelos de linguagem para domínios específicos: por exemplo, um BERT-base especializado em português jurídico via pré-treino adicional alcançou maior F1 global no LeNER-B, ligeiramente superior ao BERT-base genérico. Os experimentos confirmam essa tendência de que modelos maiores e/ou mais especializados tendem a melhor desempenho, mas também destacam os desafios do overfitting.

O contraste entre os modelos base e large merece atenção. No caso do BERT-large, foram observados ganhos modestos porém consistentes sobre o BERT-base em quase todas as classes (aumentos de aproximadamente 1–3 pontos percentuais em F1). Esse incremento, embora não dramático, indica que o modelo de maior capacidade conseguiu captar melhor algumas variações de entidade.

Em contrapartida, o RoBERTa-large trouxe poucas melhorias sobre o RoBERTa-base, na verdade. Acredita-se que isso se deve principalmente ao tamanho limitado do conjunto de treinamento. Em resumo, há um trade-off entre capacidade e dados disponíveis: modelos maiores têm mais potencial representativo, porém exigem mais dados anotados para realizá-lo plenamente. Uma lição aprendida é que, na falta de dados adicionais, pode ser mais seguro optar por um modelo menor que generaliza melhor, ao invés de um muito complexo que memorize padrões irrelevantes.

Outro ponto de discussão são os desafios específicos do domínio jurídico revelados pelos erros dos modelos. Como visto, a categoria de LOCAL foi problemática, algo atribuído à sua baixa frequência e ambiguidade no contexto legal. Identificar corretamente nomes geográficos requer contexto – um nome de cidade pode também ser sobrenome de pessoa, por exemplo – e os modelos às vezes confundiram essas classes.

Já as classes LEGISLAÇÃO e JURISPRUDÊNCIA envolvem identificar referências a normas e casos legais, que muitas vezes aparecem em formatos padronizados. Os modelos foram bastante eficazes nisso, mas ainda ocorreram confusões, especialmente confundir uma citação de caso jurídico como legislação ou vice-versa.

Em termos de aplicação real, o BERT-large desponta como o modelo mais promissor para uso no domínio jurídico dentre os testados, entregando excelente desempenho especialmente em entidades cruciais como pessoas, tempos e organizações. No entanto, do ponto de vista prático, o BERT-base pode ser uma alternativa bastante viável: sua performance ficou a apenas 3 pontos percentuais do modelo large, mas com tamanho e tempo de inferência muito menores, o que facilita a implantação em sistemas reais com recursos limitados. Dessa forma, em cenários de inferência real (por exemplo, via API), o BERT-base tende a ser mais rápido que o BERT-large, além de exigir menos memória e processamento.

Para muitas aplicações jurídicas que demandam respostas rápidas e alta disponibilidade, como assistentes jurídicos, análise de documentos em tempo real ou sistemas embarcados em tribunais, esse trade-off torna o BERT-base uma opção mais prática. Assim, a escolha do modelo ideal deve pon-

derar não apenas a acurácia bruta, mas também fatores como latência, escalabilidade e custo operacional, considerando que uma pequena perda em F1 pode ser aceitável frente aos ganhos de eficiência e usabilidade no mundo real. Em resumo, em um cenário de produção, talvez o BERT-base seja preferível se a ligeira perda de desempenho for aceitável em troca de maior velocidade e menor custo computacional.

Por fim, a engenharia da aplicação web demonstrou a usabilidade da solução. A utilização do Gradio permitiu criar rapidamente uma interface amigável, onde é possível digitar ou colar um texto e obter os resultados do NER em segundos. Isso agrega valor ao projeto, pois transforma o modelo treinado em uma ferramenta acessível a juristas e outros usuários finais.

Durante o desenvolvimento do web app, foram tomadas decisões de formatação (como destacar cada categoria com uma cor e ícone específicos) para melhorar a experiência do usuário. Em resumo, a aplicação prática mostrou que um modelo de NER jurídico pode ser integrado em interfaces e contribuir para fluxos de trabalho jurídicos, desde a análise automática de petições até a organização de base de dados de jurisprudência.

VI. Conclusão

Este trabalho apresentou uma avaliação de modelos Transformers aplicados ao reconhecimento de entidades nomeadas no domínio jurídico em português. Foram treinados e comparados os modelos: BERT-base, BERT-large, RoBERTa-base e RoBERTa-large no corpus LeNER-Br, analisando seu desempenho quantitativo e qualitativo. Verificamos que os modelos BERT alcançaram os melhores resultados, com destaque para o BERT-large, que obteve o maior $F_{1,\text{global}} \sim 90\%$.

Entretanto, considerando o custo computacional, o BERT-base mostrou-se quase tão eficaz (F1 88,6%) e significativamente mais leve, emergindo como uma opção muito promissora para uso prático em aplicações jurídicas. Os modelos RoBERTa, especialmente o large, não superaram os demais, indicando que arquiteturas multilíngues podem precisar de mais dados ou ajuste fino para atingir todo seu potencial no contexto de língua portuguesa jurídica.

Do ponto de vista de aprendizado e desenvolvimento, foi identificado a importância de equilibrar capacidade do modelo e tamanho dos dados para evitar sobreajuste. Também foram constatados desafios específicos, como a detecção de locais e a diferenciação entre legislações e casos jurídicos, que apontam para possíveis trabalhos futuros.

Pode-se concluir que os Transformers, quando devidamente adaptados, são ferramentas poderosas para NER no domínio jurídico, sendo capazes de extrair informações valiosas de textos legais com alta precisão. O projeto consolidou conhecimentos tanto em PLN jurídico quanto em desenvolvimento de software interativo. Em futuros trabalhos, a combinação de modelos maiores com corpos expandidos e o foco em casos de erro poderão alavancar o desempenho a patamares ainda mais altos, contribuindo para soluções de Inteligência Artificial no Direito cada vez mais eficazes e confiáveis.

VII. Referências

[1]Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. Lener-br: a dataset for named entity recognition in brazilian legal text. In: SPRINGER. Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018.

[2]Devlin, J., Chang, M.W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

[3]Hugging Face – Transformers Documentation. Disponível em: https://huggingface.co/transformers . Acesso em: 05 março de 2025.

[4]Kenton, J. D. M.-W. C.; Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, 2019.

[5]Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

[6]Môro, D. K. Reconhecimento de Entidades Nomeadas em Documentos de Língua Portuguesa. Dissertação de Mestrado - Universidade Federal de Santa Catarina, 2018.

[7]Souza, F., Nogueira, R., and Lotufo, R. Bert models for brazilian portuguese: Pre-training, evaluation and tokenization analysis. Applied Soft Computing, 2023.