Modelos de linguagem neurais para linguagem jurídica

Semar Augusto da C. M. Martins[1], **Adriano Veloso**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG

Abstract. *O sistema jurídico brasileiro é extremamente complexo. Possui centenas de milhares de leis e, apesar disso, é assumido que todos os cidadãos têm conhecimento delas. O auxilio de sistemas computacionais seria extremamente útil para lidar com a enorme quantidade de leis, normas, processos, decisões jurídicas e relacionados. Atualmente, tarefas de processamento de linguagem natural (PNL) dependem de uma boa representação das palavras (embeddings) para funcionar bem. A ideia desse trabalho é desenvolver um modelo de linguagem que consiga uma boa representação dos termos técnicos usados nos textos jurídicos para que esse modelo seja usado como base de sistemas de PNL que usem textos jurídicos*

1. Introduçao

O sistema jurídico brasileiro é muito complexo. Há centenas de milhares de leis e elas são editadas frequentemente. Assume-se que todo cidadão tem conhecimento da lei, mas a linguagem usada é extremamente técnica e de difícil leitura.

Palavras usadas no cotidiano da língua podem ter significados diferentes ao serem usadas no linguajar jurídico e até mesmo o contexto da escrita é diferente. Dessa maneira, um modelo de linguagem criado para a língua portuguesa de forma geral não seria suficientemente bom na hora de ser usado para tarefas que usam textos jurídicos.

O desenvolvimento de sistemas de processamento de linguagem natural (PNL), atualmente, usa o modelo de linguagem como uma forma de entrada para o modelo da tarefa a ser executada. Isso significa que a primeira etapa para o desenvolvimento de sistemas de PNL na área jurídica é o desenvolvimento de um bom modelo de linguagem para o tipo de texto a ser usado na tarefa.

Nesse trabalho demos o primeiro passo, desenvolvendo esse modelo de linguagem e disponibilizando os pesos dele.

2. Contextualização

Nos países que adotam o sistema *Common Law* tais como Estados Unidos e Inglaterra, a evolução do direito é permeada pela jurisprudência, de tal forma que os tribunais decidem promover a evolução e atualização do direito independentemente da edição de leis, apenas com a jurisprudência. Nesse sistema, numa síntese apertada, um julgamento anterior vincula próximos julgamentos.

O Brasil, por outro lado, adota o sistema do direito positivo, que, ao contrário do sistema anterior, pressupõe a criação de leis para regular as relações jurídicas. A evolução do direito depende, portanto, da constante edição e reedição das normas positivas. Nesse sistema, não existe crime sem lei anterior que o defina, mesmo que a conduta seja socialmente reprovável.

A primeira conclusão que podemos inferir disso é que, no Brasil, estamos diante de um sistema extremamente complexo e impulsionado pela edição numerosa e constante de normas positivas.

Para complicar esse aspecto, a Constituição da República dispõe, em seu artigo 5o inciso II, que, ”Ninguém será obrigado a fazer ou deixar de fazer algo senão em virtude de lei”. Donde vem o princípio segundo o qual ninguém pode alegar em sua defesa, o desconhecimento da lei. A presunção relativa, mas indispensável à manutenção da integralidade do sistema, é a de que todos os cidadãos têm conhecimento de todas as leis editadas no país. ”Presume-se que todo cidadão acorda e lê o diário oficial da União, do Estado e do Município em que reside.”

Outro aspecto que convêm destacar é que o processo legislativo de criação ou edição uma lei é, por óbvio, um processo demorado por todas as burocracias exigidas. Essa demora trava o sistema, impedindo respostas imediatas para demandas e conflitos que surgem no cotidiano. No nosso sistema, primeiro surge a demanda social e, somente depois, inicia-se o processo de legislação. Nesse hiato, o sistema decidiu colher, no sistema Common Law, soluções para regular os conflitos na ausência de leis. A solução encontrada foi a de transferir através de diferentes tipo processuais ao judiciário a regulação caso a caso dos conflitos.

Dado esse passo, e atraídos ainda por soluções eficazes do sistema *Common Law* o direito brasileiro adotou, legalmente, aspectos relativos à sumula vinculante, precedentes, paradigmas jurisdicionais que cumprem no Brasil função similar à jurisprudência no sistema *Common Law*. Em termos de prestação jurisdicional, é possível encontrar seus defensores. Contudo, isso deixa o sistema ainda mais complexo.

A partir disso, percebemos que o problema de analisar um sistema com criação e edição constante de leis, elas escritas com uma linguagem técnica e de difícil acesso à população não é uma tarefa simples. O número de leis no brasil está na casa das centenas de milhares de leis, isso sem contar as constantes edições das leis. Manter um sistema tão complexo, íntegro, é uma tarefa quase impossível.

Ao que se acresce que, embora a composição do poder legislativo não exija pessoas da área jurídica, a linguagem adotada na legislação é extremamente técnica e de difícil leitura e entendimento até mesmo para falantes nativos da lingua portuguesa.

Por isso, o desenvolvimento de sistemas computacionais capazes de analisar um grande número de leis e tomar decisões a partir disso seria extremamente útil. A primeira etapa para qualquer sistema atual baseado em processamento de linguagem natural é o desenvolvimento de um modelo de linguagem para o tipo de corpus a ser usado.

3. Bases de dados

3.1. Wikipédia em português

Foram baixados todos os artigos da Wikipédia em português com pelo menos 1800 caracteres. A decisão de fazer o filtro em 1800 caracteres foi feita para remover mensagens de erro relativas à inexistencia da pagina. Essas mensagens seriam muito frequentes no modelo de linguagem e atrapalhariam a generalização dele.

3.2. Propostas de medidas provisórias

Foram baixadas 21955 propostas de Medida Provisória (MP) do site da câmara. Elas estavam em formato PDF, das quais somente 873 estavam digitalizadas. Todas as outras 21082 propostas estavam escaneadas e, por isso, foram colocadas como entrada em um OCR (tesseract) para que obtivéssemos o texto da proposta. As emendas foram tokenizadas para remover palavras muito incomuns no corpus. Haviam muitas "palavras"no texto que só apareciam uma vez no corpus. Isso acontece devido a erros no reconhecimento da palavra pelo OCR.

3.3. Base de propostas de medidas provisórias

Utilizamos também um dataset disponível em [analytics ufcg 2019]. Esse dataset compõe 30854 linhas com informações sobre as MP's baixadas. Dentre elas há data da proposta, autor da proposta e outras informações. Foram retiradas somente duas colunas do dataset. A primeira delas continha os links para as propostas e a segunda continha a informação se as propostas foi aceitas ou rejeitadas. Essas duas colunas foram colocadas em um dataframe Pandas, de forma que o link foi usado para baixar o texto da proposta, e a informação se ela foi aceita ou não será usada para desenvolver um classificador para validar o modelo de linguagem desenvolvido no trabalho.

Das 30854 propostas no dataframe, somente foi possível fazer o download de 21955 propostas, as restantes foram ignoradas. Além disso, 608 propostas estavam com labels inválidos. Elas também foram removidas. Dentre as 21347 propostas restantes, 82.79% delas são rejeitadas.

Esse dataframe foi divido em três - treino, validação e teste. Sendo que a base de teste foi selecionada de maneira a deixar as classes perfeitamente balanceadas. Foram selecionadas 1000 propostas para a base de teste, das quais 500 foram aceitas e 500 foram rejeitadas. Essa seleção quebra a hipótese comum em aprendizado de máquina de que as bases de treino teste e validação vêm de uma mesma distribuição e são retirados de forma independente. A intenção dessa seleção é deixar a tarefa mais difícil para o modelo e garantir que ele não está prevendo que todas as propostas estão sendo rejeitadas.

4. Metodologia

O objetivo desse trabalho é é desenvolver um modelo capaz de calcular a probabilidade de uma sequência de palavras.

$$P(W) = P(w_n, w_{n-1}, ..., w_1)$$

Essa tarefa requer um conhecimento considerável da linguagem do corpus, pois é necessário entender como uma sequência de palavras é formada na língua e quais palavras ficam próximas das outras.

A arquitetura escolhida para fazer esse cálculo foi uma AWD-LSTM [Merity et al. 2017], como demonstrado na figura 1.

Esse modelo tem como tarefa calcular $P(w_n|w_{n-1}...w_1)$. Essa tarefa é equivalente à anterior pois

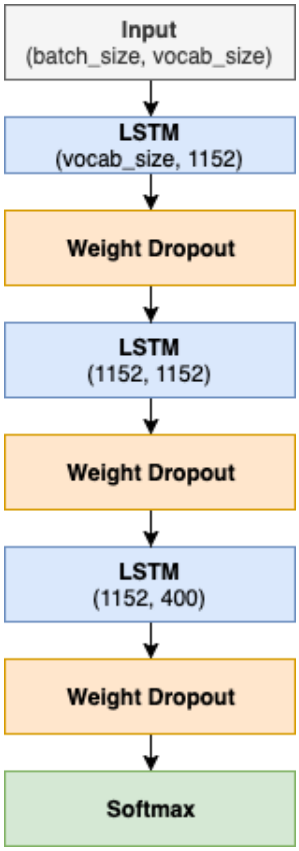

Figura 1. Arquitetura da AWD-LSTM

$$P(W) = P(w_n, w_{n-1}, ..., w_1)$$

$$P(w_n, w_{n-1}, ..., w_1) = P(w_n | w_{n-1}, ..., w_1) \times P(w_{n-1}, ...w_1)$$

Sabemos a probabilidade $P(w_{n-1}, ..., w_1)$ pois ela foi foi calculada na iteração interior do algoritmo, uma vez que o modelo consome as palavras uma a uma.

O modelo de linguagem foi treinado de acordo com o recomendado em [Howard and Ruder 2018] e ilustrado em 2. Foi desenvolvido um classificador que preveja se uma proposta de MP será aceita ou não a partir do texto escrito.

Figura 2. Metodologia usada no trabalho

Primeiro o modelo é treinado em toda a Wikipédia escrita em português. Uma vez que o modelo tem algum entendimento da lingua portuguesa, entender textos jurídicos se tornará uma tarefa muito mais fácil do que entender textos jurídicos a partir do nada. Em sequência, fazemos um ajuste fino do modelo no corpus específico que temos e por fim usamos esse modelo de linguagem como entrada para o classificador.

Todas as épocas do treinamento foram feitas usando a politica de um ciclo [Smith 2018] em conjunto com *slanted triangular learning rates* como descrito em [Howard and Ruder 2018].

Além disso, foram usadas taxas de aprendizado discriminativas [Howard and Ruder 2018] em todas as três etapas do treinamento e descongelamento gradual [Howard and Ruder 2018] no ajuste fino do modelo e no treinamento do classificador. Essas técnicas são eficiente pois diferentes camadas da rede neural tem informações de granularidades diferentes [Yosinski et al. 2014]. Camadas anteriores da rede neural aprendem estruturas mais simples do que as posteriores [Zeiler and Fergus 2013]. Assim, a mudança nos pesos das camadas anterioes deve ser menor do que a mudança nas camadas posteriores.

Em conjunto, essas técnicas deixam o treinamento mais rápido pois permitem que o modelo seja treinado usando uma taxa de aprendizado máxima maior do que o esperado, isso diminui o número de épocas necessárias para treinar o modelo.

A validação do modelo em uma tarefa de classificação é importante, pois não existe nenhum modelo de linguagem específico para linguagem jurídica que possamos usar para comparar as métricas. O classificador usado possui uma arquitetura muito simples. São somente duas camadas de perceptrons com dropout e batch normalization. A qualidade da classificação é dependente principalmente da qualidade do modelo de linguagem uma vez que o classificador sozinho seria muito limitado.

5. Resultados

5.1. Modelo de linguagem geral

O modelo de linguagem treinado na Wikipédia foi treinado por somente dez épocas, não foi possível treinar por mais tempo devido à quantidade de dados. Cada época estava demorando por volta de 5 horas em uma Titan Xp. O decorrer do treinamento pode ser visto na figura 3. Ao final foi atingido uma perplexidade de 24.655 4 uma acurácia na previsão da próxima palavra de 38.68% 5

5.2. ajuste fino do modelo de linguagem

O ajuste fino foi feito por treze épocas. O andamento do treinamento pode ser visto na figura 6, atingindo 8.029 de perplexidade (7 e 57.09% de acurácia na previsão da próxima palavra 8

Para testar se o modelo está escrevendo textos jurídicos de forma coerente, decidimos escolher uma medida provisória de 2019 (MPV 900/2019 proposta em 18/10/2019) a frase ”Fica a União, por intermédio do Ministério do Meio Ambiente, autorizada a contratar instituição financeira oficial, dispensada a licitação”para o modelo e usamos *beam search* para que ele a completasse com mais 100 palavras. Aqui estão exemplos de frases completadas pelo modelo.

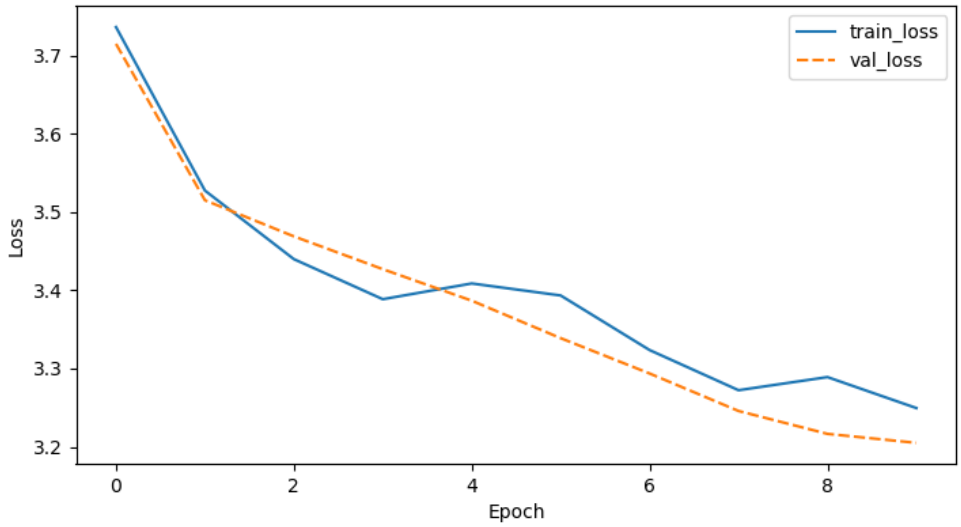

Figura 3. Função de perda ao longo do treinamento do modelo de linguagem na Wikipédia.

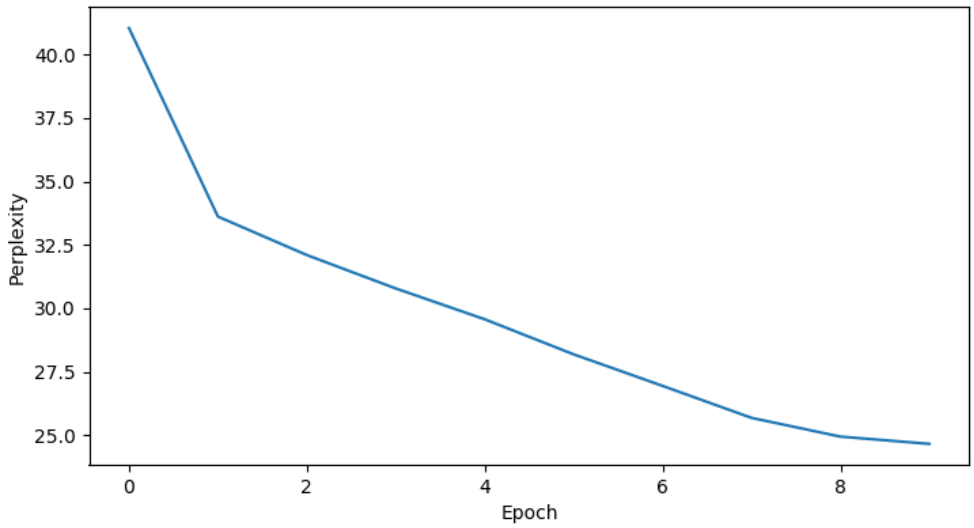

Figura 4. Perplexidade do modelo ao longo do treinamento do modelo de linguagem na Wikipédia.

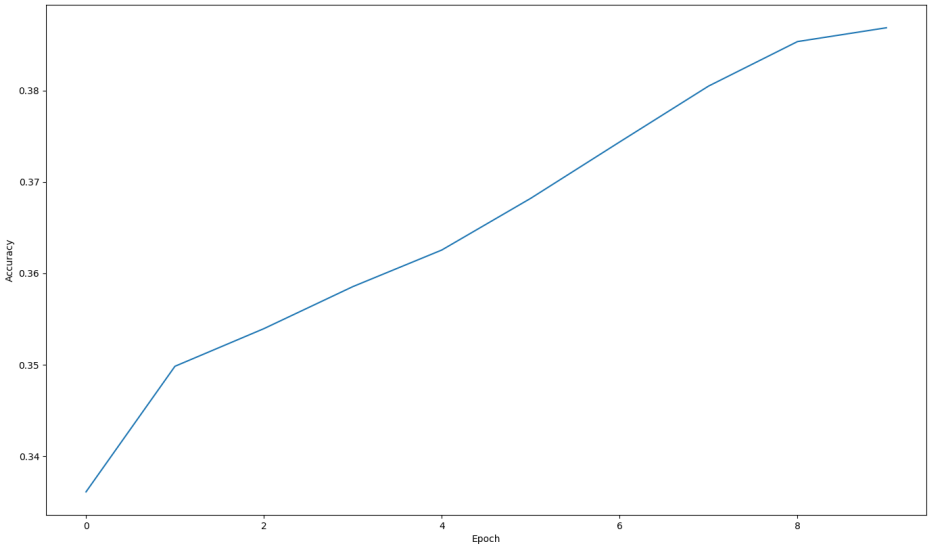

Figura 5. Acurácia do modelo ao longo do treinamento de linguagem na Wikipédia.

”Fica a União , por intermédio do Ministério do Meio Ambiente , autorizada a contratar instituição financeira oficial , dispensada a licitação .

Art . Fica autorizada a adoção das seguintes medidas de estímulo a liquidação ou à NEGOCIAÇÃO de dívidas originárias de operações de crédito rural inscritas na DAU ou que venham a ser incluídas até 30 de novembro de 2014 :

i — concessão de descontos , conforme quadro constante do Anexo IX desta Lei , para a liquidação da dívida até 30 de dezembro de 2014 , devendo incidir o desconto percentual sobre a soma dos saldos devedores por mutuário na data da renegociação , observado o disposto no 5 10 deste artigo , e , em seguida , ser aplicado o respectivo desconto de valor fixo por faixa de saldo devedor ;

Il — permissão da renegociação do total dos saldos devedores das operações até 30 de dezembro de 2014 ,

”Fica a União, por intermédio do Ministério do Meio Ambiente, autorizada a contratar instituição financeira oficial, dispensada a licitação Fica a União , por intermédio do Ministério do Meio Ambiente , autorizada a contratar instituição financeira oficial , dispensada a licitação .

Art . Fica autorizada a adoção das seguintes medidas de estímulo a liquidação ou à renegociação de dívidas originárias de operações de crédito rural inscritas na DAU ou que venham a ser incluídas até 30 de novembro de 2013 :

i - concessão de descontos , conforme quadro constante do Anexo IX desta Lei , para a liquidação da dívida até 30 de dezembro de 2014 , devendo incidir o desconto percentual sobre a soma dos saldos devedores por mutuário na data da renegociação , observado o disposto no 5 10 deste artigo , e , em seguida , ser aplicado o respectivo

desconto de valor fixo por faixa de saldo devedor ;

ll — permissão da renegociação do total dos saldos devedores das operações até 30 de dezembro de 2014 , mantendo

Testando com a seguinte frase, retirada da medida provisória Nº 893, de 19 de agosto 2019 "Transforma o Conselho de Controle de Atividades Financeiras na Unidade de Inteligência Financeira. O PRESIDENTE DA REPÚBLICA, no uso da atribuição que lhe confere o art. 62 da Constituição, adota a seguinte Medida Provisória, com força de lei Transforma o Conselho de Controle de Atividades Financeiras na Unidade de Inteligência Financeira . o PRESIDENTE DA REPÚBLICA , no uso da atribuição que lhe confere o art . 62 da Constituição , adota a seguinte Medida Provisória , com força de lei"obtivemos:

"Transforma o Conselho de Controle de Atividades Financeiras na Unidade de Inteligência Financeira. O PRESIDENTE DA REPÚBLICA, no uso da atribuição que lhe confere o art. 62 da Constituição, adota a seguinte Medida Provisória, com força de lei Transforma o Conselho de Controle de Atividades Financeiras na Unidade de Inteligência Financeira . o PRESIDENTE DA REPÚBLICA , no uso da atribuição que lhe confere o art . 62 da Constituição , adota a seguinte Medida Provisória , com força de lei

Art . 1° a Carreira de que trata esta Lei é composta do cargo de Policial Rodoviário Federal , estruturada nas classes de Inspetor , Agente Especial , Agente e Inicial , na forma do Anexo i desta Lei .
5 1ª São requisitos para o ingresso na carreira o diploma de curso superior completo , em nível de graduação , devidamente reconhecido pelo Ministério da Educação , bem como os demais requisitos estabelecidos no edital do concurso .

5 2º a investidura no cargo de Policial Rodoviário Federal dar - se - á no padrão único da classe Inicial , onde permanecerá por , pelo menos , três anos ou até obter o direito à promoção à

"Transforma o Conselho de Controle de Atividades Financeiras na Unidade de Inteligência Financeira. O PRESIDENTE DA REPÚBLICA, no uso da atribuição que lhe confere o art. 62 da Constituição, adota a seguinte Medida Provisória, com força de lei Transforma o Conselho de Controle de Atividades Financeiras na Unidade de Inteligência Financeira . o PRESIDENTE DA REPÚBLICA , no uso da atribuição que lhe confere o art . 62 da Constituição , adota a seguinte Medida Provisória , com força de lei :

Art . 1º a Carreira de que trata esta Lei é composta do cargo de Policial Rodoviário Federal , estruturada nas classes de Inspetor , Agente Especial , Agente e Inicial , na forma do Anexo i desta Lei .

5 1º As atribuições gerais das classes do cargo de Policial Rodoviário Federal são as seguintes :

i - classe de Inspetor : atividades de natureza policial e administrativa , envolvendo direção , planejamento , coordenação , supervisão , controle e avaliação administrativa e operacional , coordenação e direção das atividades de corregedoria , bem como a articulação e o intercâmbio com outras organizações e corporações policiais , em âmbito nacional e internacional , além das atribuições da classe

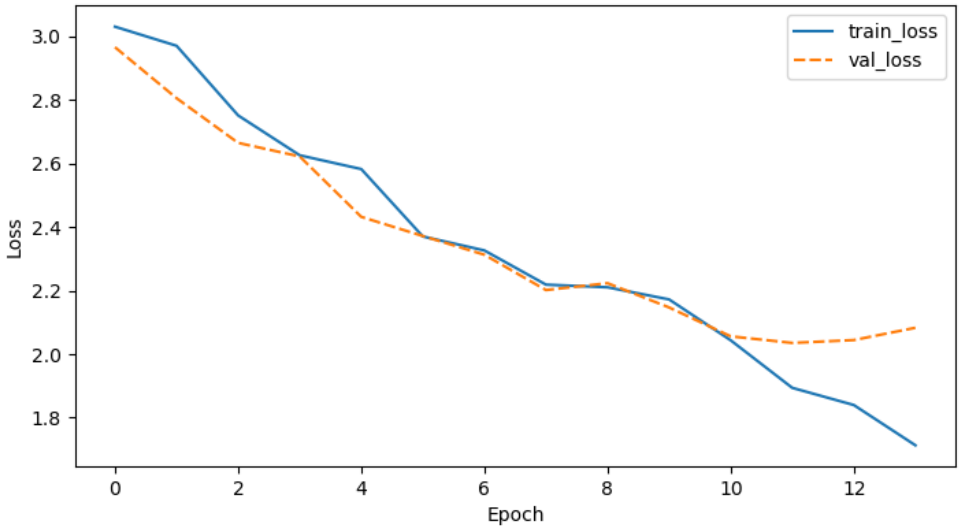

Figura 6. Função de perda ao longo do ajuste fino do modelo de linguagem.

Notamos que a linguagem usada está coerente, apesar do texto em si não fazer sentido, como é esperado de textos gerados por modelos de linguagem. Há um problema no modelo que veio devido à forma que os dados de propostas forma obtidos. O Tesseract substitui que o símbolo §, comumente usado nos textos das propostas de lei pelo número 5. Por isso o modelo de linguagem começa algumas frases com o número 5.

5.3. Classificador

O classificador foi treinado por um total de 8 épocas, o decorrer do treinamento pode ser visto em 9. A seleção de hiper-parâmetros foi feita usando um otimizador bayesiano [Snoek et al. 2012] com o objetivo de maximizar o f1-score do modelo. O otimizador cria uma distribuição de funções a posteriori da função alvo e, à medida que o número de observações aumenta, começa a determinar quais espaços da distribuição merecem ser explorados. O otimizador fez a seleção da taxa de aprendizado, *weight decay* e *dropout*.

Devido à seleção da base de treino com diferente distribuição, é interessante mostrar tanto os resultados na base de validação a figura 10 demonstra como ocorreu a variação as métricas de acurácia e f1-score do modelo avaliadas na base de validação ao final de cada época. Podemos notar que ao final do treinamento foi obtida uma acurácia de 90.09% e um f1-score de 0.9410 na base de validação. Ela reflete melhor a distribuição real do dado e portanto deve ser levada em consideração como a provável acurácia real do modelo.

A base de teste, por sua vez, conseguiu uma acurácia de 88.9% e um f1-score de 0.9395. Podemos ver na matriz de confusão 11 que o modelo tem uma performance melhor nas propostas que são rejeitadas do que nas propostas que são aceitas. Isso pode ser facilmente explicado pela distribuição dos dados, uma vez que 82% das propostas são de fato rejeitadas. Assim, o modelo teve acesso a muito mais dado que explica o que faz uma proposta ser rejeitada do que dado que explica o que faz uma proposta ser aceita.

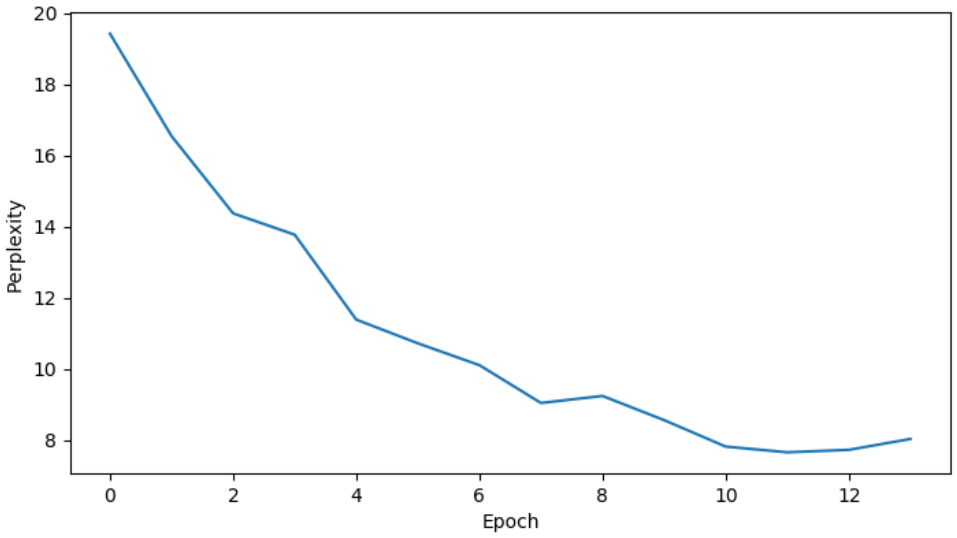

Figura 7. Perplexidade do modelo ao longo do ajuste fino do modelo de linguagem.

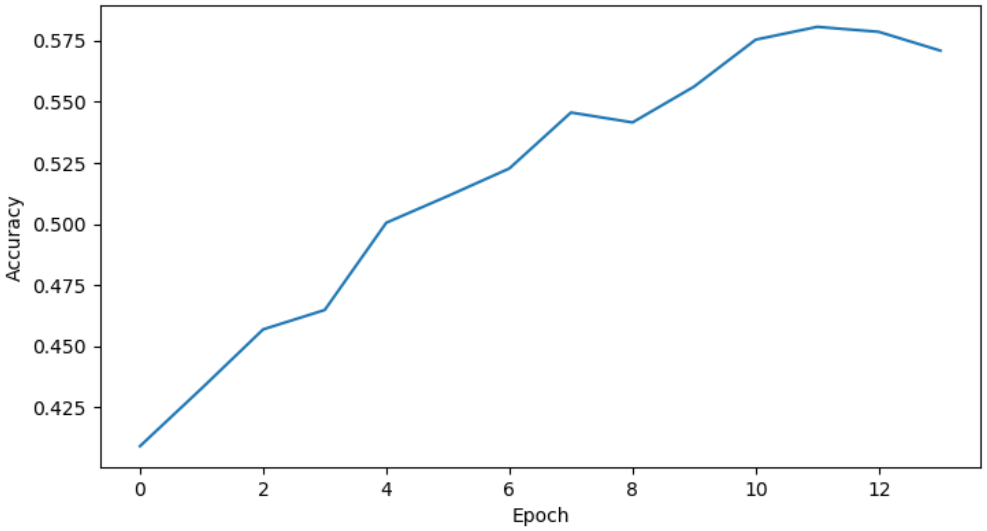

Figura 8. Acurácia do modelo ao longo do ajuste fino do modelo de linguagem.

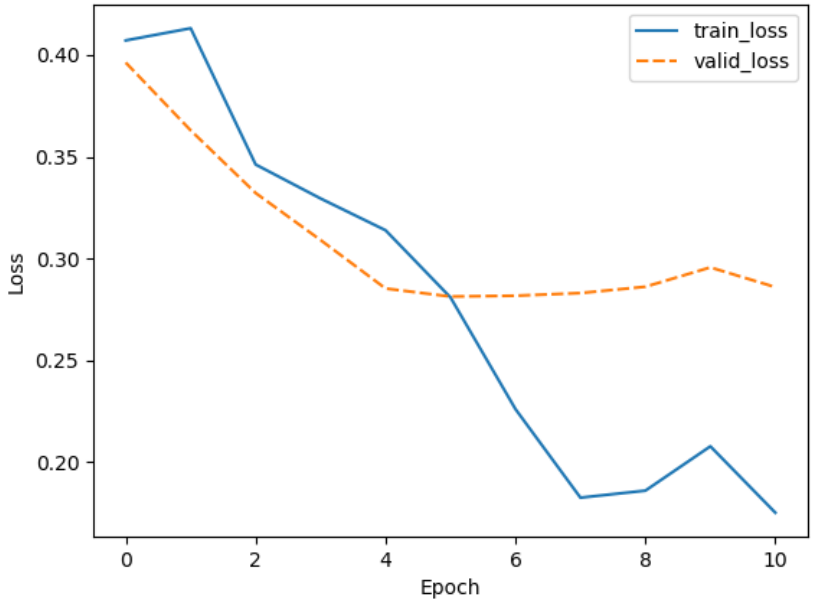

Figura 9. Função de perda média avaliada na base de treinamento e validação

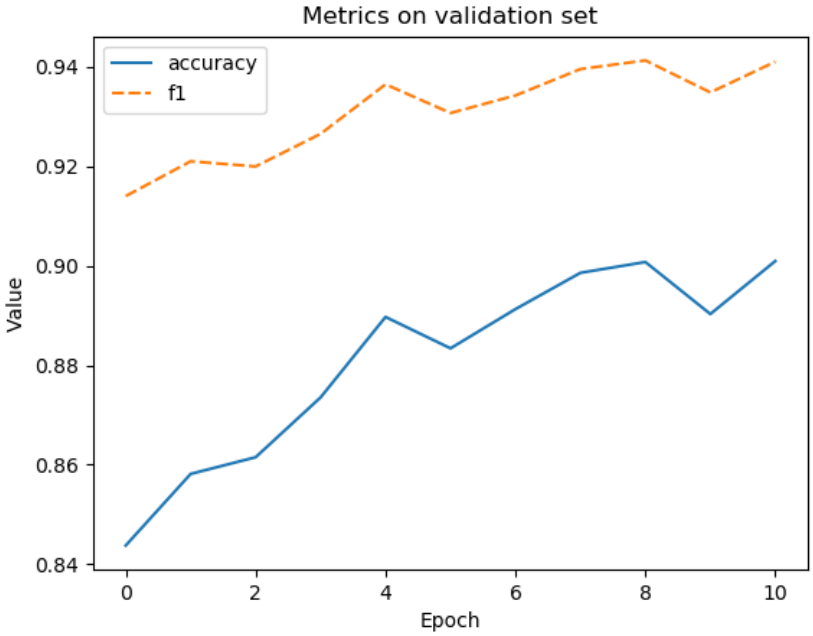

Figura 10. Acurácia e F1-Score medidoas na base de validação ao fim de cada época

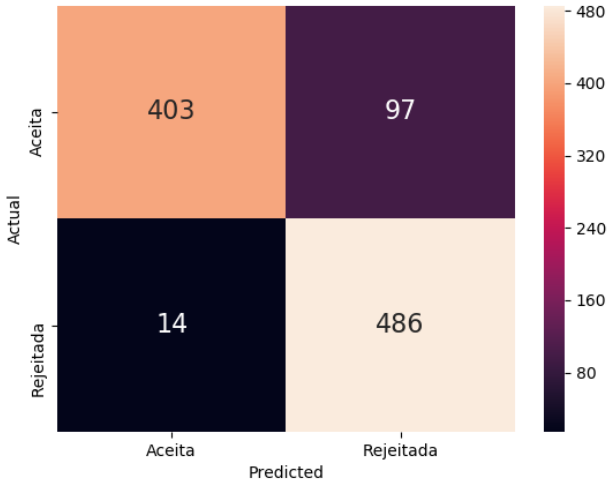

Figura 11. Matriz de confusão do modelo avaliada na base de teste.

6. Conclusão

Nesse trabalho desenvolvemos um modelo de linguagem para textos jurídicos usando uma arquitetura AWD-LSTM e, devido à falta de baseline, fizemos um classificador que depende do modelo de linguagem. O classificador prediz se uma proposta de medida provisória será ou não aceita na câmara.

Ainda é inconclusivo determinar se o classificador poderia ser usado na prática devido ao possível viés existente na base de dados usada durante o trabalho. Seria necessário testar o modelo com uma amostra aleatória de propostas de períodos diferentes.

Referências

analytics ufcg (2019). Projeto leg(islativo) go do ddc/ufmg. `https://github.com/analytics-ufcg/leggo-content/tree/crawler_igor`.

Howard, J. and Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

Merity, S., Keskar, N. S., and Socher, R. (2017). Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR*, abs/1803.09820.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *CoRR*, abs/1411.1792.

Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.