Avaliação de modelos de DeepLearning para Reconhecimento de Entidades Nomeadas no Domínio Legal em Língua Portuguesa

Hidelberg O. Albuquerque[1]
1Centro de Informática, Universidade Federal de Pernambuco, Recife, Brasil
Email: hoa@cin.ufpe.br

***Abstract*— Lorem ipsum**

Keywords—xx

I. INTRODUÇÃO

As tarefas de Reconhecimento de Entidades Nomeadas (NER) são um desafio no Processamento de Linguagem Natural para a língua portuguesa. Tarefas de NER podem ser focadas na identificação de categorias semânticas dentro do texto, como Pessoas, Localização, Organizações entre outras. Outras tipologias semânticas também poder desenvolvidas para representar necessidades específicas do domínio estudado. O domínio jurídico ou legal inclui uma grande variedade de textos, como leis, processos legais, acórdãos, diários oficiais, entre outros documentos. Existem trabalhos relacionados ao domínio legal e tarefas de NER em outras línguas, contudo, devido à particularidade dos modelos e documentos em língua portuguesa, acredita-se ser necessário o desenvolvimento de abordagens e ferramentas de alta qualidade para extração de informação como ocorre em outras línguas..

II. TRABALHOS RELACIONADOS

Lorem ipsum.

III. PROPOSTA

No desenvolvimento desta pesquisa foi executado um benchmark a partir da experimentação do desempenho de modelos de Deep Learning aplicados ao estado da arte, aplicados à dois corpora do domínio legal: Lener-Br Corpus (artigo base) e PL-Corpus (corpora desenvolvido a partir de Projetos de Lei da Câmara dos Deputados Brasileira).

Utilizando cunho exploratório, esta pesquisa foi executado através de uma metodologia experimental, inicialmente aplicando o modelo de aprendizado Bi-LSTM+CRF aos corpora selecionados, buscando reproduzir o estado da arte. Após esta fase inicial, foram aplicadas otimizações nos hiperparâmetros do modelo buscando melhoria do desempenho. Num segundo momento, foi utilizada a arquitetura Transformers (utilizando BERT), buscando novamente a otimização dos hiperparâmetros. Em ambos os casos, foram utilizadas as métricas Precisão, Recall e F1-score para avaliação nas tarefas de NER em textos legais.

IV. EXPERIMENTOS

Inicialmente, para se ter um objeto de comparação, foi aplicado o código do artigo-base[1] utilizando o modelo Bi-LSTM+CRF e o corpus disponibilizados pelos autores, sem alterações nos parâmetros originais. Não foi encontrado no artigo-base a quantidade de execuções efetuadas e, para manter a simetria com o método utilizado neste trabalho, foram feitas cinco execuções e calculadas a média e desvio-padrão geral e por entidade. A Tabela I apresenta os resultados encontrados, com destaque em negrito para os melhores scores. Como demonstrado, de maneira geral, os resultados obtidos pelos autores do artigo-base foram superiores aos encontrados na aplicação deste trabalho, o que possibilitou esta investigação.

Desta forma, buscando melhoramos nos resultados inicialmente obtidos, foram executados os experimentos como descriminado a seguir:

- Modelo Bi-LSTM+CRF: a) avaliação de modelos de vetores de palavras (*word embbedings*) na etapa de pré-processamento; b) otimização do modelo através dos hiperparâmetros *batch size*, *épocas* e *métodos de otimização*;
- Modelo Transformers (BERT): a) utilização de um modelo pré-treinado para domínio legislativo, com fine-tunning com o artigo base; b) otimização do modelo através dos hiperparâmetros: *batch size*, épocas e taxas de aprendizado.

O código-fonte do modelo Bi-LSTM+CRF e o corpus utilizado no artigo base está disponível no github dos autores[2]. O modelo BERT foi adaptado de [Guilhou, 2021][3]. Os corpora foram divididos conjunto de treinamento (75%) e teste (25%). Para a execução dos experimentos utilizando otimização de hiperparâmetros, foram feitas analises isoladas por grupos e pela combinação dos hiperparâmetros selecionados. Para cada grupo de hiperparâmetros foram feitas cinco execuções, modificando aleatoriamente os conjuntos de treinamento e teste, e calculado a média e desvio padrão. Para a etapa de permutação de hiperparâmetros, os conjuntos de treinamento e teste foram randomizados uma única vez. Os experimentos foram executados no Google Colab, utilizando a infraestrutura de hardware GPU NVidia-SMI 460.32.03, com 27.3 GB de memória RAM.

Buscando estabelecer um link entre corpora de mesmo domínio, também aplicamos os experimentos acima ao PL-Corpus. Os códigos modificados e os corpora utilizados estão disponíveis na *github* do autor deste trabalho[4].

A. Bi-LSTM+CRF

Vetores de Palavras (word embbedings) são representações numéricas de cada palavra (*tokens*) do corpus e o grau de similaridade com as demais palavras do corpus.

[1] *https://www.researchgate.net/publication/327751993_LeNER-Br_A_Dataset_for_Named_Entity_Recognition_in_Brazilian_Legal_Text*

[2] *https://github.com/peluz/lener-br*

[3] *https://huggingface.co/pierreguillou/ner-bert-base-cased-pt-lenerbr*

[4] *https://github.com/bergoliveira/disciplinaDL*

TABLE I.
COMPARAÇÃO COM LENER-BR (ARTIGO-BASE)

Entidades	Lener-Br (artigo base)			Lener-Br (aplicação neste trabalho)		
	Precisão	***Recall***	***F1-Score***	***Precisão***	***Recall***	***F1-Score***
Pessoa	**94.44%**	92.52%	93.47%	94.13±5.31%	**93.66±8.17%**	**93.85±7.04%**
Local	**61.24%**	**59.85%**	**60.54%**	45.41±12.24%	55.81±14.37%	49.39±11.61%
Organização	**91.27%**	**85.66%**	**88.38%**	83.59±6.63%	74.09±14.19%	78.07±12.05%
Tempo	**91.15%**	**91.15%**	**91.15%**	86.56±1.81%	88.51±3.06%	87.49±1.81%
Legislação	**97.08%**	**97.00%**	**97.04%**	86.49±4.62%	85.85±8.8%	86.05±7.21%
Jurisprudência	**87.39%**	**90.30%**	**88.82%**	68.45±10.86%	64.92±12.87%	66.24±12.07%
Overall	**93.21%**	**91.91%**	**92.53%**	82.33±3.33%	78.53±10.18%	80.12±7.89%

Originalmente, o código-base utiliza o vetor de palavras *Glove*, treinado para o português brasileiro. Existem outros modelos de *word embbedings* pré-treinados e disponíveis para acesso público[5]. Foram utilizados na avaliação, além do Glove, os vetores Word2Vec, Wang2Vec e FastText, e avaliados o impacto destes vetores. Em todos os casos, foi aplicado os modelos com 300 dimensões.

Após esta etapa, foi investigada o impacto da Otimização de Hiperparâmetros no modelo de aprendizagem, visando otimizar melhorar o resultado inicial através da combinação de um conjunto de variáveis pré-definidas. O espaço de configurações possível pode ser arbitrário ou aleatório. Baseando-se nos parâmetros originais do artigo-base, e por questões de infraestrutura, arbitrariamente foram escolhidas os grupos/combinações demonstradas na Tabela II. Os valores destacados são os valores originais do código-base. Além da análise por grupos, foi feita também a combinação de todos os parâmetros, e escolhida o que obteve melhor resultado na média.

TABLE II.
HIPERPARÂMETROS DE TREINAMENTO BI=LSTM+CRF

Table Column Head	Valores
Batchs (ambos)	[**10**, 20, 40]
Épocas (ambos)	[35,**55**,75]
Método de otimização (Bi-LSTM+CRF)	[**SGD**, ADAM, ADAGRAD, RMSProp]
Taxa de aprendizado (BERT)	[1e-4, 2e-5, 3e-4, 5e-5]

B. Transformers (com BERT)

O modelo BERT é um modelo de deep learning pré-treinado de código aberto para Transfomers. Foi utilizado o modelo *NER model in the legal domain in Portuguese (LeNER-Br)* [Guilhou, 2021], seguido da etapa de fine-tunning utilizando o dataset do Lener-Br[6]. O treinamento deste modelo possibilitou um resultado interessante, que serviu para investigar o refinamento de hiperparâmetros do modelo, como demonstrado na Tabela III. Assim como executado no treinamento da rede Bi-LSTM+CRF, as inicialmente foram analisados o impacto de modificação de grupos de parâmetros, executando posteriormente a combinação de possibilidades. Os valores destacados na Tabela representam os valores originais do modelo Bi-LSTM+CRF, aplicados no modelo BERT. Vale uma observação: não foi possível observar no modelo BERT um parâmetro associado à métodos de otimização e, por isso, foi decidido utilizar os parâmetros comentados no código do modelo.

V. RESULTADOS E DISCUSSÕES

A. Bi-LSTM+CRF

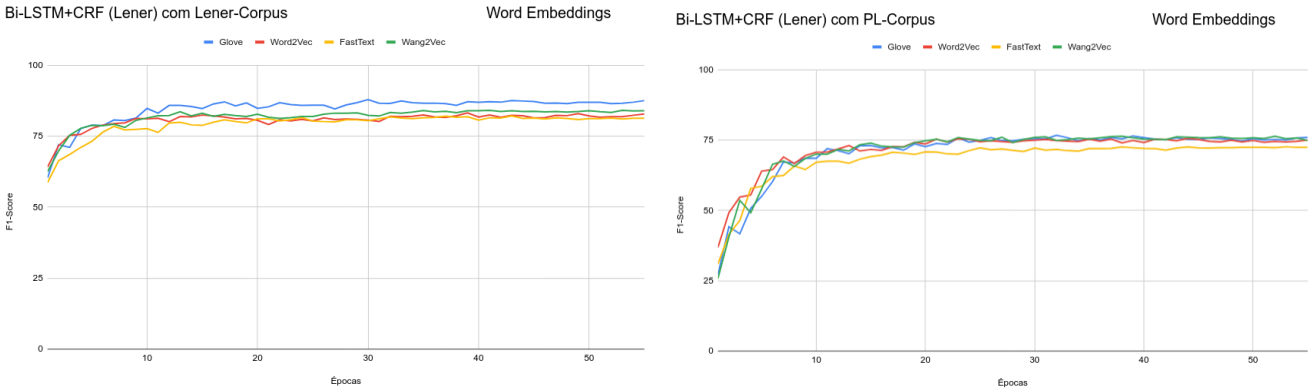
Como demonstrado na Fig. 1, o modelo de *word embbedings* que melhor se destacou foi o *Glove* quando aplicado no corpus Lener-Br, replicando o resultado do artigo-base. Devido sua melhor performance, foi escolhido este vetor de palavras como padrão para os demais testes neste modelo de aprendizagem. Vale a pena destacar um comportamento comum a todos os vetores aplicados: ao chegar numa época em foi alcançado o platô de melhor F1-score, este platô médio de score se mantinha, ocorrendo a partir da época 30 em diante, em ambos os corpora. À exceção a este comportamento foram os vetores FastText e Wang2Vec, que só vieram alcançar este patamar no final do treinamento do PL-Corpus.

Analisando agora os grupos de hiperparâmetros, para cada grupo testado foram mantidas as demais configurações originais do modelo: (i) buscou-se investigar inicialmente o efeito do tamanho do batch na etapa de treinamento e o intervalo de generalização. Por questões de instraturura de hardware, iniciou-se o texto com o mesmo valor do modelo-base[7], dobrando seu valore a cada teste. Como demonstrado na Fig. 2(a-b), o aumento do tamanho batch para o mesmo número de épocas não trouxe benefícios para o treinamento, mas um decréscimo médio na precisão na tarefa de NER, em ambos os corpora (Fig. 2c); (ii) analisando o impacto no número de épocas nos conjuntos de treinamento dos corpora selecionados, é possível observar uma pequena tendência de melhoramento no treinamento, principalmente quando aplicado o modelo ao PL-Corpus (Fig. 2d-f); (iii) ainda utilizando os mesmos hiperparâmetros e modificando os mé-

[5] *http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc*

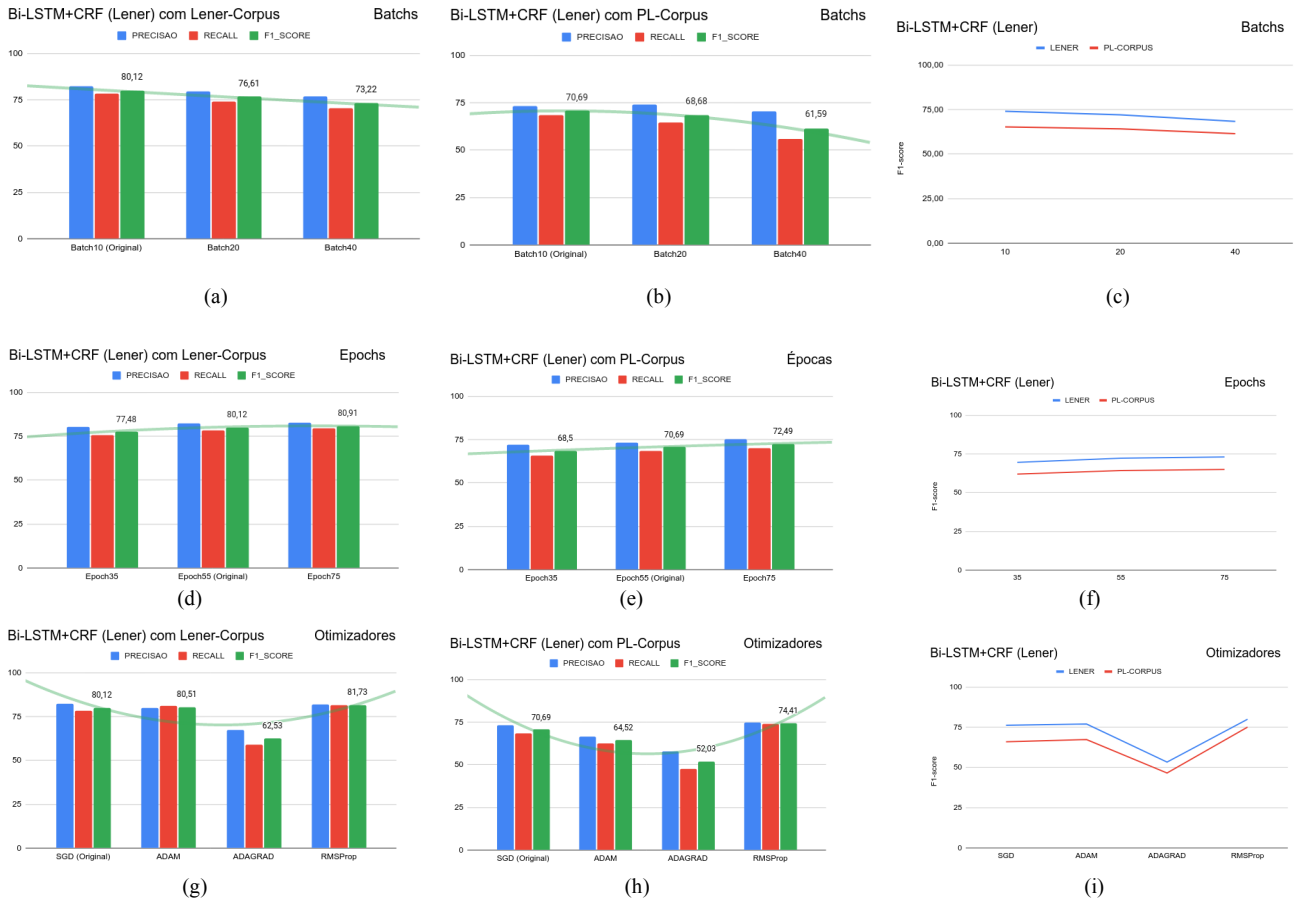
[6] *https://huggingface.co/datasets/lener_br*

[7] Nas testes a seguir, o termo "original" é utilizado para referenciar os parâmetros originais do modelo, mas com os valores obtidos no experimento inicial obtido neste trabalho, demonstrado na Tabela I.



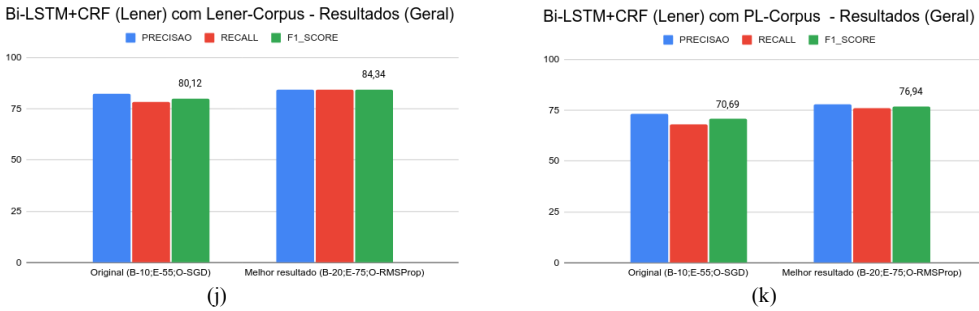
(a) (b)

Fig. 1. Analise de Word Embbedings no treinamento do modelo Bi-LSTM+CRF: (a) aplicando o corpus Lener, melhor resultado com o Glove, demais vetores se mantém na média. (b) Pior resultado com o modelo FastText no PL-Corpus, e os demais modelos na média.



(d) (e) (f)

(g) (h) (i)



(j) (k)

Fig. 2. Bi-LSTM+CRF: analise de hiperparâmetros por grupo: (a-c) não há influência positiva aumentando o número de Batchs; (d-f): maior número de épocas demonstrou tendência de melhoramento; (g-i): método de otimização com melhor resultado foi RMSProp;

todos de otimização, o modelo original vem pré-configurado com o otimizador SVD. Nos testes utilizados, destaca-se um leve ganho dos otmizadores ADAM e RMSProp na tarefas de NER para os corpus utilizados, valendo a pena destacar o desempenho muito baixo do ADAGRAD em todos os casos (vide Fig. 2g-i); por fim, buscando encontrar uma melhor configuração para as tarefas de NER no conjunto de treinamento nos corpora utilizados, foi executada a permutação de todas as possibilidades no conjunto de dados, incluindo na análise as já feitas anteriormente (por batchs, épocas e otimizadores). Assim, como demonstra a Fig. 2j e Fig. 2k, houve uma sincronia no treinamento, encontrando o melhor conjunto de hiperparâmetros, para ambos os corpora treinados: Batchs: 20 + Épocas: 75 + Otimizador: RMSProp. Destacamos um aumento de F1-score do conjunto originalmente treinado nesta pesquisa de 5,2% e 8,8% nos corpora Lener e PL-Corpus, respectivamente. Observamos mediante os dados apresentados anteriormente que a probabilidade de a quantidade de épocas associado ao otimizador gradiente RMSProp terem sido os causadores deste ganho. As Tabela III e IV demonstra este ganho à nível de F1-score, por entidade em cada corpus.

TABLE III.
BI-LSTMF+CRF COM LENER CORPUS: F1-SCORES POR ENTIDADES

	F1-score	
Entidade	**Anterior**	**Novo**
Pessoa	93.85	**97.55**
Local	49.39	**61.29**
Organização	78.07	**79.94**
Tempo	87.49	**88.42**
Legislação	86.05	**86.15**
Jurisprudência	66.24	**86.27**
Overall	80.12	**84.34**

TABLE IV.
BI-LSTMF+CRF COM PL-CORPUS: F1-SCORES POR ENTIDADES

	F1-score	
Entidade	**Anterior**	**Novo**
Pessoa	72.65	**79.34**
Local	72.97	**79.5**
Organização	60.15	**73.01**
Data	88.86	**89.53**
Fundamento	75.98	**79.44**
Produto de Lei	32.28	**42.55**
Overall	70.69	**76.94**

B. Transformers (com BERT)

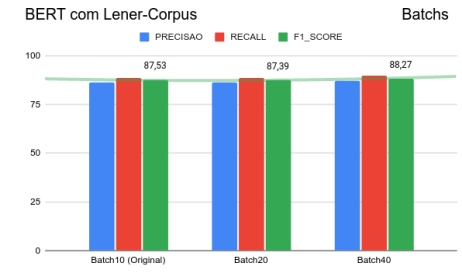
Buscando avaliar os modelos pré-treinados com BERT para o domínio legal, com fine-tunning com o dataset Lener-Br, optou-se por seguir a mesma metodologia executada na avaliação na seção anterior. Como o modelo BERT não utiliza uma etapa de pré-processamento de texto, esta etapa não foi utilizada. Como dito anteriormente, foi treinado um modelo utilizando os parâmetros de batchs e épocas iguais ao Bi-LSTM+CRF, e iniciando a taxa de aprendizado = 2e-5, como arbitrado no modelo pré-treinado de origem.

Iniciando a análise dos resultados (i) pelo tamanho do batch na etapa de treinamento, é possível perceber pela Fig. demonstrado na Fig. 3(a-b), uma leve tendência de aumento pontual e médio. Todavia, à medida que se aumentava o tamanho dos batchs, a etapa de treinamento ficava utilizava mais espaço em disco, o que, na nossa avaliação, não justificaria o seu uso; (ii) de forma similar, na análise do aumento do número de épocas (Fig. 3c-d), não houve um aumento pontual nas tarefas de NER, mas chama a atenção a queda encontrada na média de aprendizado para a época 55, muito provavelmente ocasionada pela combinação com a taxa de aprendizagem ruim; (iii) quanto à taxa de aprendizagem (Fig. 3e-f), o índice = 3e-4 apresentou baixíssima qualidade, em momentos chegando a não identificar nenhuma entidade no treinamento, mas destacando um aumento interessante quando aplicado o índice de aprendizagem = 5e-5; (iv) por fim, após a combinação dos hiperparâmetros utilizados no conjunto de treinamento para o corpus Lener-Br (Fig. 3-g), pode ser verificado que o conjunto com os parâmetros que melhor resultaram foi Batchs: 40 + Épocas: 55 + Taxa de aprendizado: 2e-5. A Tabela V apresenta os resultados por entidade. Até o presente, os resultados da aplicação destes conjuntos para o PL-Corpus não estavam completos para serem apresentados.

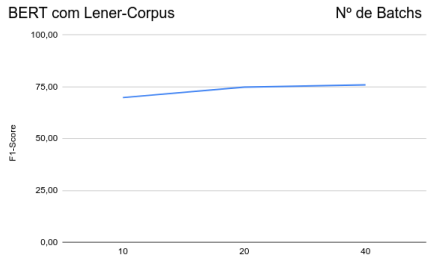
TABLE V.
BERT COM LENER CORPUS: F1-SCORES POR ENTIDADES

	F1-score	
Entidade	**Anterior**	**Novo**
Pessoa	**98,89**	97,93
Local	81,36	**83,33**
Organização	**89,17**	88,98
Tempo	96,99	**97,63**
Legislação	85,82	**86,22**
Jurisprudência	74,88	**77,72**
Overall	87,53	**88,27**

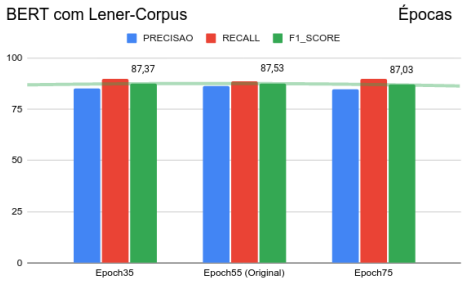
Por fim, a Fig. é possível verificar que, entre os resultados obtidos nos modelos testados neste trabalho, o modelo pré-treinado BERT obteve melhor score quando aplicado às tarefas de NER para o Lener-Corpus, obtendo F1-Score médio de 88,27%, contra o melhor resultado encontrado com Bi-LSTM+CRF de 84.34%. Todavia, ainsa são necessárias mais investigações para descobrir como alcançar os resultados obtidos no artigo-base. Além disso, ainda falta a completude da análise do PL-Corpus utilizando o modelo BERT, para comparação.



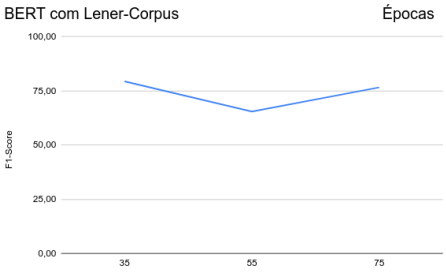
(a)



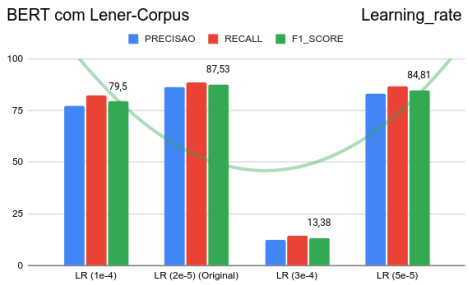
(b)



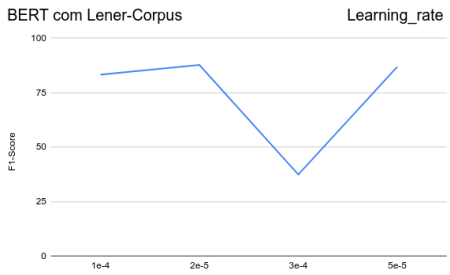
(c)



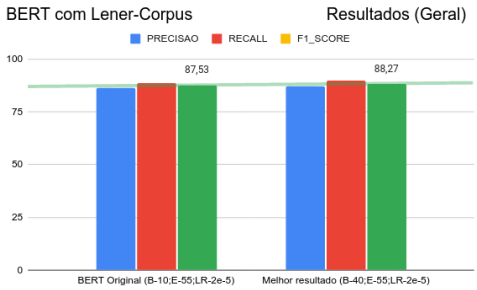
(d)



(e)



(f)



(g)

Fig. 3. BERT: analise de hiperparâmetros por grupo (a-b): pequena tendência de crescimento médio pelo aumento de Batchs; (c-d): praticamente sem modificação no F1-score; (e-f): influência na taxa de aprendizado, com destaque para o desempenho ruim de 3e-4; (g) resultado final.

ACKNOWLEDGMENT *(Heading 5)*

Todo.

REFERENCES

Todo.