gilBERTo: UM MODELO DE LINGUAGEM PARA A LÍNGUA PORTUGUESA

ARTHUR ALVES CERVEIRA[1]; LARISSA ASTROGILDO DE FREITAS[2]; ULISSES BRISOLARA CORRÊA[3]

[1]UNIVERSIDADE FEDERAL DE PELOTAS – aacerveira@inf.ufpel.edu.br
[2]UNIVERSIDADE FEDERAL DE PELOTAS – larissa@inf.ufpel.edu.br
[3]UNIVERSIDADE FEDERAL DE PELOTAS – ulisses@inf.ufpel.edu.br

1. INTRODUÇÃO

A interação entre humanos e máquinas através da linguagem natural é um dos maiores desafios dos campos da computação e da linguística, e a área do conhecimento responsável por estudar essa interação é chamada de Processamento de Língua Natural (PLN). Uma abordagem para problemas de PLN que se tornou muito popular nos últimos anos é a utilização de Modelos de Linguagem (ML), que são modelos treinados para calcular a probabilidade de uma determinada sequência de palavras ocorrer em uma frase (JURAFSKY, 2020). Os ML são capazes de generalizar as propriedades e regras do idioma em que foram treinados, e podem ser utilizados como base para realizar tarefas mais específicas na área de PLN. A popularidade desses modelos pode ser explicada em parte pela utilização de textos não estruturados para seu treinamento, que são amplamente disponíveis em diversos domínios e idiomas.

Recentemente, abordagens baseadas em redes neurais se mostraram como soluções mais robustas para o treinamento de ML. Em 2019, foi publicada a metodologia BERT para o treinamento de ML, com base na arquitetura *Transformer* (DEVLIN, 2019). Essa arquitetura introduz o mecanismo de atenção no treinamento de redes neurais, que é capaz de melhor assimilar as informações contextuais do que arquiteturas mais tradicionais (VASWANI, 2017). Esse mecanismo também permite um maior paralelismo do que arquiteturas anteriores, reduzindo de forma significativa seu tempo de treinamento. Os modelos baseados BERT (e outros modelos derivados dessa arquitetura) apresentam hoje soluções estado-da-arte para diversos problemas trabalhados na área de PLN (ROGERS, 2020), e já estão presentes em ferramentas difundidas entre o público geral.

Os ML possuem grande potencial de inovação acadêmica em razão da sua capacidade de generalizar regras de sintaxe e semântica a partir de grandes conjuntos de textos não estruturados. Esses modelos podem ser então utilizados como base para realizar tarefas mais específicas (RUDER, 2019). O procedimento de treinar um ML para essa finalidade é denominado pré-treinamento. A partir dessa abordagem, modelos mais simples podem tirar proveito da representação de linguagem aprendida por um modelo pré-treinado, e ajustá-lo para atingir um objetivo específico. A utilização desses modelos pré-treinados tem como grande vantagem reduzir a quantidade de dados rotulados necessários em tarefas que envolvem o aprendizado supervisionado. Esses dados rotulados costumam ser extremamente escassos em idiomas como o Português. Portanto, o pré-treinamento de modelos baseados em BERT para a língua portuguesa são essenciais para o avanço da área de PLN neste idioma.

Sendo assim, este trabalho tem como objetivo geral pré-treinar e avaliar um novo modelo de linguagem baseado em BERT para o idioma Português. O resultado esperado ao final deste trabalho é possuir o modelo pré-treinado

disponível para a comunidade acadêmica, o qual pode ser utilizado em diversas tarefas de PLN. Para atingir o objetivo geral, foram definidos como subprodutos um conjunto de objetivos específicos que devem ser cumpridos. Dentre os objetivos específicos, podemos citar: levantar corpora não estruturados que serão utilizados no pré-treinamento; gerar o vocabulário do modelo; pré-treinar o modelo; identificar quais são as soluções estado-da-arte para as principais tarefas realizadas na área de PLN; avaliar o modelo pré treinado nessas tarefas; e entender qual o impacto da utilização dessas técnicas mais modernas no desempenho do modelo.

2. METODOLOGIA

Esta seção descreve os métodos de pré-treinamento e avaliação do modelo proposto, apelidado de gilBERTo neste trabalho.

A metodologia de pré-treinamento do modelo inicia na etapa de construção do conjunto de dados, que será composto de *corpora* não-estruturados de domínio geral. Esses dados passarão por um processo de limpeza, que visa reduzir ruídos que podem estar presentes nos textos, e eliminar características indesejadas. Parte desse conjunto de dados será então utilizado como entrada para a construção do vocabulário do modelo, através da utilização de um *tokenizer* para segmentar as palavras e sub-palavras mais frequentes em *tokens*. Ambos o conjunto de dados e o vocabulário serão utilizados na criação dos dados de pré-treinamento. Por fim, esses dados de pré-treinamento, em conjunto a um *checkpoint* de um modelo BERT pré-treinado e um arquivo de configuração, serão utilizados como entrada para executar o pré-treinamento do modelo. O resultado esperado é ter como saída um modelo BERT pré-treinado para o Português. Esse processo de pré-treinamento é apresentado na Figura 1.

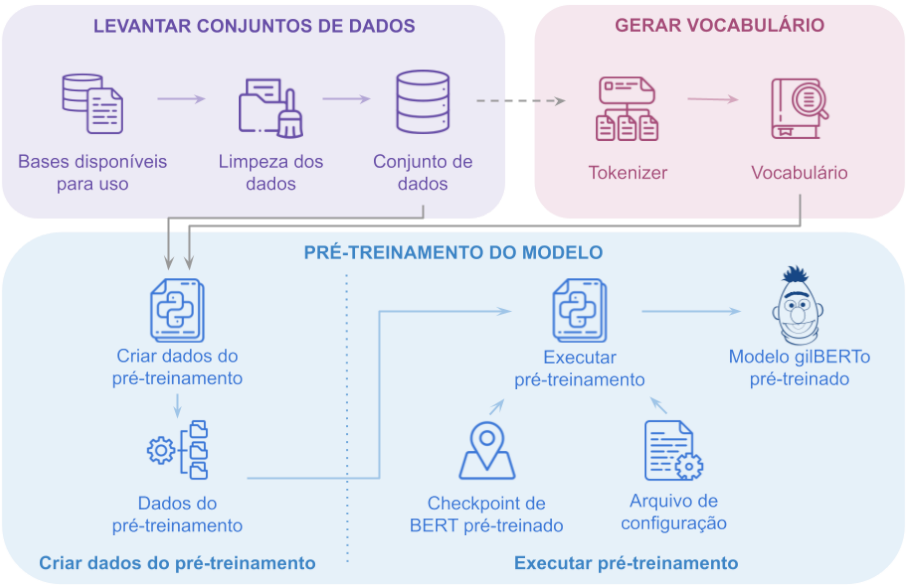

Figura 1: Metodologia de pré-treinamento do modelo de linguagem gilBERTo para o Português. Fonte: Própria.

A metodologia de avaliação do modelo parte de uma revisão bibliográfica, onde serão selecionadas as tarefas e conjuntos de avaliação mais relevantes da área de PLN para o Português. Para a comparação dos resultados obtidos nessas

tarefas, serão identificados outros ML estado-da-arte no Português e multi-língua que serão utilizados como linha de base. Para realizar esse processo, será necessário implementar códigos de ajuste e aplicação do modelo nos conjuntos de teste, ou então reutilizar códigos que foram disponibilizados em trabalhos anteriores. Todos os resultados serão compilados e analisados, fornecendo uma visão de como o modelo treinado neste trabalho se compara com as demais soluções da literatura. Para a avaliação do modelo, foi considerada a tarefa de Reconhecimento de Entidades Nomeadas (REN).

3. RESULTADOS E DISCUSSÃO

Nesta seção serão apresentadas as três tarefas consideradas para a avaliação e comparação do modelo proposto com outros ML baseados em BERT. Foram considerados os modelos M-BERT, BERTimbau e BertPT para comparação com o modelo proposto. A seguir serão definidos os objetivos da tarefa, os conjuntos de dados, e as métricas de avaliação. Foi avaliada a abordagem *fine-tuning* para treinar os modelos na execução das tarefas alvo.

A tarefa REN tem como objetivo reconhecer as entidades nomeadas em um texto através de um processo de classificação de *tokens* em uma sequência. Os conjuntos de dados contém textos de domínios diversos, e as entidades nomeadas foram classificadas em 10 categorias: Pessoa, Organização, Local, Valor, Data, Título, Coisa, Evento, Abstração e Outro. A Tabela 1 apresenta os resultados obtidos pelos modelos através da abordagem fine-tuning, considerando o uso dos modelos com e o classificador CRF para reconhecer as entidades nomeadas. A principal para avaliação dessa tarefa é a F1, enquanto as métricas de Precisão e Revocação são auxiliares. Nesse contexto, a métrica de Precisão representa a porcentagem de entidades nomeadas pelo modelo que estão corretas, a Revocação é a porcentagem de entidades presentes no corpus que foram corretamente preditas, e o F1 é então calculado a partir da média harmônica entre essas duas métricas. Foi utilizado o conjunto de dados *FirstHAREM* para o treinamento e *MiniHAREM* para avaliação dos modelos.

Tabela 1: Comparação entre modelos BERT treinados na tarefa REN

Modelo	Precisão	Revocação	F1
M-BERT	0,73	0,7254	0,7277
BERTimbau	0,7973	0,7817	0,7894
BertPT	0,2632	0,2397	0,2509
gilBERTo	0,7618	0,7342	0,7478

Nessa comparação, é possível observar que o modelo BERTimbau atingiu os melhores resultados na métrica principal da tarefa no conjunto de dados utilizado para avaliação. O modelo gilBERTo apresentou um desempenho competitivo, atingindo resultados superiores aos do M-BERT nas 3 métricas consideradas. Em todos os cenários considerados, o BertPT atingiu resultados inferiores aos demais modelos avaliados. Para os modelos BertPT e BERTimbau, esses resultados são condizentes com o desempenho constatado pelos autores nos trabalhos que introduziram esses modelos. No geral, os resultados para as métricas de Precisão e Revocação se mantiveram similares aos resultados da

métrica principal de avaliação. Esse resultado é esperado, visto que essas duas métricas são utilizadas no cálculo do F1.

4. CONCLUSÕES

Este trabalho apresenta o pré-treinamento e avaliação do gilBERTo, um ML baseado na metodologia BERT para a língua portuguesa. A metodologia deste trabalho possui como diferencial em relação a outros modelos da literatura o uso de técnicas mais modernas de treinamento de ML, desde um processo mais rigoroso de limpeza dos textos de treinamento e do vocabulário, até as tarefas de pré-treinamento e implementação que foram consideradas. Foi utilizado também o conjunto de dados OSCAR que é frequentemente adotado para o treinamento de ML em outras línguas, mas não havia sido utilizado em modelos BERT para o Português até o presente momento. Entre os resultados obtidos, o modelo se manteve próximo ao desempenho atingido pelo M-BERT, não superando a performance do modelo BERTimbau nas tarefas escolhidas como *benchmark*, ainda que atinja resultados superiores aos do BertPT. Levando em conta a duração reduzida do seu tempo de treinamento em relação a outros modelos avaliados, o modelo proposto ainda foi capaz de atingir resultados competitivos nas tarefas consideradas. Este trabalho tem como pontos fortes o desenvolvimento de uma metodologia transparente de pré-treinamento e avaliação, com a disponibilização dos códigos e o modelo que poderão ser utilizados e melhorados pela comunidade de PLN. Ainda assim, este trabalho teve como limitação os recursos computacionais disponíveis para a execução da metodologia, que são essenciais para viabilizar o pré-treinamento de grandes ML. Os resultados e artefatos atingidos neste trabalho apresentam potencial de continuidade dessa pesquisa no prosseguimento do pré-treinamento do gilBERTo, no treinamento de ML maiores e com arquiteturas mais modernas, e também a aplicação do modelo pré-treinado em outras tarefas da área de PLN.

5. REFERÊNCIAS BIBLIOGRÁFICAS

JURAFSKY, D. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Stanford, CA, 2020. 3v.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **Association for Computational Linguistics**, Minneapolis, v.1, n.15, p.4171-4186, 2019.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; AND JONES, L.; GOMEZ, A., LUKASZ, I. Attention is all you need. **Advances in neural information processing systems**, Long Beach, v.30, n.15, 2017.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A Primer in BERTology: What We Know About How BERT Works. **Transactions of the Association for Computational Linguistics**, Cambridge, v.8, n.14, p.842-866, 2020.

RUDER, S.; PETERS, M.; SWAYAMDIPTA, S.; WOLF, T. Transfer Learning in Natural Language Processing. **Association for Computational Linguistics**, Minneapolis, v.1, n.3, p.15-18, 2019.