Comparando Modelos Compactos e Especializados para Análise de Sentimento em Tweets

Leandro I. Freitas[1] **, Anselmo C. Paiva**[1] **Fael Faray de Paiva** [1]
Arnaud Guedes de Paiva Neto [1]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
Caixa Postal: 65.085-580 – São Luís, MA – Brasil

{leandro.iglesias, anselmo.paiva, Fael.paiva, Arnauld.Páiva}@nca.ufma.br

Abstract. *High-performance sentiment analysis is now dominated by large language models, whose high computational cost imposes a critical trade-off between performance and efficiency. This work investigates this relationship by empirically comparing representative models from two optimization strategies—model compression (DistilBERT, ALBERT) and domain specialization (BERTweet)—on the task of tweet sentiment classification. The results demonstrate that while domain specialization achieves the highest predictive performance, compressed models, particularly DistilBERT, present a highly competitive cost-benefit balance. This study, therefore, offers a quantitative guide for selecting architectures in practical scenarios, elucidating the relative advantages of each approach.*

Resumo. *A análise de sentimentos de alta performance é hoje dominada por grandes modelos de linguagem, cujo alto custo computacional impõe um trade-off crítico entre desempenho e eficiência. Este trabalho investiga essa relação ao comparar empiricamente modelos representativos de duas estratégias de otimização — compressão (DistilBERT, ALBERT) e especialização de domínio (BERTweet) — na tarefa de classificação de sentimentos em tweets. Os resultados demonstram que, embora a especialização de domínio alcance o maior desempenho preditivo, os modelos compactos, especialmente o DistilBERT, apresentam um balanço custo-benefício altamente competitivo. Este estudo, portanto, oferece um guia quantitativo para a seleção de arquiteturas em cenários práticos, elucidando as vantagens relativas de cada abordagem.*

1. Introdução

A análise de sentimentos em mídias sociais é uma tarefa fundamental para a compreensão da opinião pública na era digital, consolidada como subárea central de *opinion mining* [Pang and Lee 2008], com aplicações que abrangem monitoramento de marca, análise política, previsão de mercado e gestão de crises [Liu 2015]. Contudo, apresenta desafios específicos decorrentes da natureza informal, ruidosa e frequentemente mal estruturada do texto nessas plataformas, marcada por gírias, emoticons, abreviações, limitações de caracteres e alta dependência de contexto [Rosenthal et al. 2017, Cambria et al. 2013]. A construção de corpora específicos para mídias sociais, como o Twitter, evidenciou a complexidade desse domínio e consolidou sua importância como ambiente de teste

para modelos de Processamento de Linguagem Natural (PLN) [Pak and Paroubek 2010, Go et al. 2009b].

Nesse cenário, modelos de linguagem pré-treinados baseados na arquitetura Transformer, como o BERT [Devlin et al. 2019a], estabeleceram um novo patamar de desempenho em tarefas de PLN, incluindo análise de sentimentos. Entretanto, o aumento contínuo no tamanho e na complexidade desses modelos impõe custos computacionais elevados para treinamento e inferência, restringindo seu uso prático em cenários com recursos limitados e levantando preocupações ambientais e éticas quanto ao consumo energético e às emissões de carbono associadas [Strubell et al. 2019, Bender et al. 2021, Patterson et al. 2021]. Em paralelo, discute-se a necessidade de arquiteturas mais eficientes, capazes de equilibrar desempenho preditivo e custo computacional [Tay et al. 2022a].

Diante desse contexto, duas estratégias complementares têm ganhado destaque: **compressão de modelos** e **especialização de domínio**. A compressão engloba técnicas como poda, esparsidade estruturada e arquiteturas compactas [Blalock et al. 2020, Frankle and Carbin 2019, Zhou et al. 2019], bem como destilação de conhecimento, em que um modelo professor maior transfere seu aprendizado a um modelo estudante menor, preservando boa parte do desempenho com redução substancial de parâmetros [Hinton et al. 2015]. O DistilBERT exemplifica essa linha ao apresentar uma versão reduzida do BERT, mais rápida e leve, mantendo desempenho competitivo em diversas tarefas [Sanh et al. 2019].

A especialização de domínio, por sua vez, adapta modelos genéricos a contextos específicos por meio de pré-treinamento contínuo ou ajuste orientado a dados de um determinado domínio, estratégia que se mostra eficaz para capturar padrões linguísticos particulares [Gururangan et al. 2020]. No contexto de mídias sociais, o BERTweet representa essa abordagem ao ser pré-treinado especificamente em tweets, superando modelos de uso geral em tarefas de classificação de sentimentos no Twitter [Nguyen et al. 2020]. Em conjunto, essas linhas de pesquisa evidenciam que tanto a compressão quanto a adaptação ao domínio são caminhos promissores para tornar modelos mais utilizáveis em cenários reais.

Entretanto, a escolha entre arquiteturas compactas de uso geral e modelos maiores porém especializados envolve um *trade-off* não trivial entre desempenho, eficiência computacional, acessibilidade tecnológica e impacto ambiental. Enquanto modelos especializados tendem a capturar melhor as particularidades do domínio, modelos comprimidos oferecem vantagens claras em tempo de inferência e uso de memória, aspectos cruciais para aplicações em larga escala ou em dispositivos com restrição de recursos [Cai et al. 2020]. Persiste, contudo, uma lacuna na literatura quanto à quantificação sistemática desses *trade-offs* em cenários concretos de análise de sentimentos em mídias sociais, nos quais o volume massivo de dados e a alta variabilidade linguística tornam as decisões de arquitetura ainda mais sensíveis.

Diante disso, este trabalho realiza uma análise comparativa entre três modelos representativos de estratégias distintas: BERTweet, como modelo especializado em tweets [Nguyen et al. 2020]; DistilBERT, como exemplo de compressão via destilação de conhecimento [Sanh et al. 2019, Hinton et al. 2015]; e ALBERT, como arquitetura extremamente leve baseada em compartilhamento de pesos e fatorização de embeddings

[Lan et al. 2020]. Utilizando o corpus Sentiment140 [Go et al. 2009b] em uma tarefa de classificação binária (positivo vs. negativo), avaliamos simultaneamente desempenho e custo computacional, considerando acurácia, precisão, *recall*, tempo de inferência e consumo de memória. O objetivo é oferecer evidências empíricas que auxiliem na escolha informada do modelo mais adequado a diferentes restrições de recursos e requisitos de aplicação, contribuindo para esclarecer, de forma integrada, o *trade-off* entre performance e eficiência na análise de sentimentos em mídias sociais.

2. Trabalhos Relacionados

A introdução da arquitetura Transformer [Vaswani et al. 2017] e dos modelos de linguagem pré-treinados [Devlin et al. 2019b] redefiniu fundamentalmente a abordagem para a análise de sentimentos em texto. O paradigma dominante que emergiu consiste em realizar um ajuste fino (*fine-tuning*) desses modelos massivos em conjuntos de dados específicos para a tarefa. A eficácia dessa abordagem é amplamente documentada, mas a escolha do modelo e da estratégia de otimização ideais permanece um campo de pesquisa ativo, especialmente ao se considerar o *trade-off* entre desempenho e custo computacional.

Nesse contexto, uma vertente da literatura foca em maximizar a performance preditiva. Trabalhos como o de [Sun et al. 2019] investigaram as melhores práticas para o ajuste fino do BERT, enquanto revisões como a de [Adu-Acheampong et al. 2021] consolidam a dominância dessas arquiteturas. Para domínios específicos como o de mídias sociais, modelos especializados como o **BERTweet** [Nguyen et al. 2020] foram propostos, demonstrando que o pré-treinamento em dados de nicho pode levar a resultados de ponta, superando modelos de uso geral.

Paralelamente, uma segunda vertente de pesquisa aborda o desafio da eficiência computacional, propondo arquiteturas ”leves”para reduzir o custo de inferência com perda mínima de acurácia [Tay et al. 2022b]. Estudos comparativos são fundamentais nessa área, como o de [Bhatt et al. 2020], que analisou a latência de variantes do BERT em benchmarks gerais. Outros, como [Ganesh et al. 2021], dissecaram a relação entre o número de parâmetros e o desempenho. Contudo, esses estudos de eficiência raramente colocam os modelos compactos em uma comparação direta com modelos altamente especializados, deixando uma lacuna na análise de domínios ruidosos e dinâmicos como o das mídias sociais.

É precisamente nesta lacuna, a comparação direta entre as filosofias de especialização para performance e de compressão para eficiência, que o presente estudo se posiciona. O trabalho se diferencia ao conduzir uma análise empírica controlada que coloca modelos representativos de ambas as estratégias (DistilBERT e ALBERT para compressão; BERTweet para especialização) em uma competição direta no desafiador domínio dos tweets. Ao quantificar não apenas as métricas de desempenho (Acurácia, F1, Precisão, Recall), mas também as de custo (Parâmetros, FLOPs, Tempo), este estudo oferece um guia empírico claro sobre qual estratégia proporciona o balanço mais favorável para aplicações práticas de análise de sentimentos em mídias sociais.

3. Metodologia

O fluxo metodológico deste estudo abrange desde a preparação do corpus até a definição das métricas de avaliação e a estratégia de treinamento dos modelos. portanto, todos os experimentos foram desenvolvidos em `Python 3.10`, utilizando as bibliotecas `PyTorch 2.2` e `HuggingFace Transformers 4.41` [Wolf et al. 2020]. A execução foi realizada em uma unidade de processamento gráfico (GPU) NVIDIA RTX 3060 com 12 GB de VRAM.

3.1. Corpus e Pré-processamento

O estudo utilizou o dataset Sentiment140, apresentado originalmente por Go, Bhayani e Huang (2009). Este corpus de grande escala, com 1,6 milhão de tweets, foi rotulado por meio de supervisão distante (distant supervision). Conforme descrito pelos autores, o processo envolveu a coleta de tweets que continham emoticons positivos para serem automaticamente rotulados como positivos (rótulo 4) .e, de forma análoga, para os tweets com emoticons negativos. Após essa atribuição automática, os emoticons foram removidos do texto. Essa abordagem cria um corpus massivo sem a necessidade de anotação manual, forçando o modelo a aprender o sentimento a partir do conteúdo semântico do texto, em vez de depender da simples presença de um emoticon [Go et al. 2009a].

Nesse viés, optou-se por utilizar apenas 2% do corpus original (aproximadamente 32.000 instâncias), com uma divisão balanceada entre as duas classes (positivo e negativa). Para o pré-processamento, os campos de metadados irrelevantes para a tarefa (user, date, id, flag) foram descartados. Os rótulos de sentimento foram mapeados para um formato binário, onde o rótulo original 4 (positivo) foi convertido para 1.

Finalmente, o subconjunto de dados foi particionado de forma estratificada em conjuntos de treino (80%), validação (10%) e teste (10%). A estratificação assegura que a distribuição de classes seja consistente em todas as partições, prevenindo vieses durante o treinamento e a avaliação.

3.2. Arquiteturas Avaliadas

Foram selecionadas três variantes do modelo BERT, cada uma representando uma abordagem distinta de otimização, seja por compactação ou por especialização de domínio:

1. **DistilBERT-base-uncased** [Sanh et al. 2019]: Um modelo obtido através da destilação de conhecimento (knowledge distillation) do BERT. Essa técnica resulta em um modelo com aproximadamente 66 milhões de parâmetros, significativamente menor e mais rápido que seu professor, mantendo grande parte de sua capacidade preditiva.
2. **ALBERT-base-v2** [Lan et al. 2020]: Uma arquitetura que emprega técnicas de compartilhamento de parâmetros entre as camadas (cross-layer parameter sharing) e fatorização da matriz de embedding para reduzir drasticamente a complexidade do modelo para cerca de 12 milhões de parâmetros, focando na eficiência paramétrica.
3. **BERTweet-base** [Nguyen et al. 2020]: Um modelo de 135 milhões de parâmetros especializado em linguagem de mídias sociais. Foi pré-treinado em um corpus massivo de 850 milhões de tweets, tornando-o intrinsecamente adaptado às particularidades lexicais e sintáticas deste domínio.

3.3. Tokenização

Para cada modelo, foi utilizado o tokenizador correspondente disponibilizado na biblioteca HuggingFace. Os textos foram padronizados para um comprimento máximo de `max_length=128` tokens, através de truncamento ou preenchimento. Este valor foi considerado suficiente para abranger a totalidade do conteúdo da grande maioria dos tweets. Conforme a recomendação dos autores de BERTweet, o parâmetro `add_prefix_space=True` foi habilitado para garantir que o primeiro token do tweet seja processado corretamente.

3.4. Estratégia de Ajuste Fino

Todos os modelos passaram por um processo de ajuste fino para a tarefa de classificação de sentimentos binária. O processo de treinamento foi gerenciado pela API `Trainer` da HuggingFace para garantir consistência e reprodutibilidade. Os hiperparâmetros foram mantidos idênticos entre os modelos para permitir uma comparação justa de desempenho, conforme detalhado abaixo:

- **Épocas de Treinamento:** 3
- **Tamanho do Lote (Batch Size):** 8 para treino e 16 para validação/teste
- **Taxa de Aprendizagem (Learning Rate):** 2×10^{-5} com otimizador AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Aquecimento (Warm-up):** 500 passos iniciais
- **Decaimento de Peso (Weight Decay):** 0,01

Adotou-se uma estratégia de early stopping implícita, onde o melhor modelo foi selecionado com base no maior valor de Macro F1-score obtido no conjunto de validação ao final do treinamento (`load_best_model_at_end=True`). O monitoramento das métricas e o registro dos experimentos foram realizados com o auxílio da plataforma Weights & Biases (W&B).

3.5. Métricas de Avaliação

O desempenho dos modelos foi quantificado utilizando um conjunto abrangente de métricas, calculadas para cada partição do corpus (treino, validação e teste) com o suporte da biblioteca `scikit-learn` [Pedregosa et al. 2011]:

- **Acurácia (Accuracy):** Percentual de predições corretas.
- **Macro F1-Score:** Média harmônica da precisão e revocação, calculada de forma independente para cada classe e depois ponderada igualmente, robusta a desbalanceamentos residuais.
- **Precisão (Precision)** e **Revocação (Recall):** Calculadas especificamente para a classe positiva (rótulo 1).
- **Custo Computacional e Complexidade:** Para avaliar a eficiência, registramos o número de parâmetros de cada modelo, o total de operações de ponto flutuante (FLOPs) para a inferência no conjunto de teste e o tempo de treinamento..

Não foi aplicada ponderação de classes (class weights), uma vez que a amostragem estratificada garantiu um balanceamento quase perfeito entre as classes.

4. Resultados e discussão

Os resultados do estudo são apresentados em duas frentes: a análise de desempenho (Tabela 1) e a de custo computacional (Tabela 2).

A Tabela 1 detalha as métricas de performance, onde o BERTweet-base demonstra clara superioridade na maioria dos quesitos, incluindo acurácia (84,47%) e precisão (85,31%). Este resultado reforça a eficácia do pré-treino em domínio específico. Embora o desempenho do DistilBERT, seja inferior em acurácia, alcançou o maior valor de recall (84,76%), indicando uma maior sensibilidade para identificar todos os exemplos positivos.

Tabela 1. Resultados de desempenho.

Modelo	Acurácia (%)	Macro-F1 (%)	Precisão (%)	Recall (%)
DistilBERT	81.28	81.98	79.38	**84.76**
ALBERT	78.84	78.60	79.94	77.30
BERTweet	**84.47**	**84.38**	**85.31**	83.46

A superioridade do BERTweet em termos de acurácia geral pode ser atribuída ao seu pré-treinamento em um corpus de 840 milhões de tweets, permitindo que o modelo capture idiossincrasias linguísticas específicas do domínio, como a ocorrência de hashtags, menções de usuários (@) e linguagem coloquial. Este achado alinha-se com a literatura que demonstra que modelos especializados apresentam ganhos consistentes de 3-5% em tarefas específicas de domínio quando comparados a modelos genéricos [Kavitha et al. 2025].

O desempenho do DistilBERT tem recall superior (84,76%) sugerindo que, apesar de ter 50% menos parâmetros que o BERTweet, o modelo mantém uma boa capacidade de generalização. Isso é particularmente relevante para aplicações onde a detecção de falsos negativos é crítica, como em sistemas de monitoramento de crise que precisam identificar todas as manifestações negativas de sentimento [Ramesh et al. 2023]. A precisão ligeiramente inferior (79,38%) indica uma taxa maior de falsos positivos, trade-off esperado e aceitável em muitos cenários práticos.

O ALBERT apresentou o desempenho mais baixo entre os três modelos (78,84% de acurácia), com uma queda de aproximadamente 5,6% em relação ao BERTweet. Esta degradação de performance sugere que a estratégia de compressão agressiva empregada pelo ALBERT (compartilhamento de pesos entre camadas e fatorização de embeddings) pode ser excessiva para capturar a complexidade linguística de tweets [Xie et al. 2024]. O modelo teve particular dificuldade em manter o recall (77,30%), indicando dificuldades em identificar instâncias positivas em dados ruidosos de Twitter.

Em contrapartida, a análise de custo e complexidade, detalhada na Tabela 2, revela o claro trade-off entre performance e eficiência.

Tabela 2. Análise de custo computacional e complexidade.

Modelo	Parâmetros (M)	FLOPs ($\times 10^{15}$)	Tempo (min)
DistilBERT	≈ 66	2.54	4.57
ALBERT	≈ 12	0.46	8.03
BERTweet	≈ 135	5.05	8.79

4.1. Análise do Custo Computacional e Eficiência

O alto desempenho do BERTweet está associado ao maior custo computacional, tanto em número de parâmetros (135 M) quanto em FLOPs ($5,05 \times 10^{15}$). Essa carga computacional adicional tem implicações significativas para implementação em ambientes com restrições de recursos, como dispositivos móveis, *edge computing* e servidores com limitações de memória. O tempo de treinamento de $8,79$ minutos representa um custo energético considerável, especialmente quando multiplicado por múltiplos experimentos ou ajustes finos em escala de produção.

O DistilBERT consolida-se como uma alternativa de excelente custo-benefício. Este modelo exigiu quase metade dos recursos computacionais do BERTweet ($2,54$ vs $5,05$ FLOPs) para atingir um desempenho robusto. Particularmente notável é o tempo de treinamento ($4,57$ minutos), aproximadamente 45% mais rápido que seus concorrentes. Para organizações operando em larga escala, essa diferença se amplifica significativamente, resultando em economias substanciais de energia e custo operacional. A abordagem de destilação de conhecimento empregada pelo DistilBERT provou ser particularmente efetiva neste contexto, retendo aproximadamente $\sim 96\%$ do desempenho de BERT base com 40% de redução paramétrica.

O ALBERT-base-v2 posiciona-se como a arquitetura mais econômica em parâmetros (apenas 12 M) e FLOPs ($0,46 \times 10^{15}$), representando uma redução de 91% em parâmetros em relação ao BERTweet. Essa eficiência extrema viabiliza *deployment* em cenários restritos, como aplicações embarcadas e processamento em tempo real em dispositivos periféricos. Porém, essa eficiência resulta em uma queda sensível nas métricas de desempenho ($78,84\%$ de acurácia), representando uma degradação de $5,6$ pontos percentuais em relação ao BERTweet. Essa relação não-linear entre compressão e perda de *performance* é bem documentada na literatura de otimização de modelos e sugere limites práticos para o quanto um modelo pode ser reduzido sem sacrifício excessivo de funcionalidade.

A análise integrada de ambas as tabelas revela uma relação entre especialização/compressão e desempenho preditivo. O DistilBERT emerge como um ponto ótimo nesse *continuum*: alcança $\sim 96\%$ do desempenho do BERTweet, consumindo apenas cerca de 50% dos recursos computacionais. Essa relação sugere que, para análise de sentimentos no Twitter, a estratégia de destilação oferece um balanço superior à compressão agressiva (ALBERT) ou à especialização com alto custo computacional (BERTweet).

A importância relativa deste *trade-off* varia significativamente conforme o contexto de aplicação. Para organizações operando em escala *massive* (processando milhões de *tweets* diariamente), a economia de tempo computacional do DistilBERT ($4,57$ vs $8,79$

minutos por época) pode resultar em redução de $40 - 50\%$ no gasto com infraestrutura de computação[Tay et al. 2022b]. Conversamente, para aplicações onde a precisão é absolutamente crítica—como análise de *feedback* de clientes em setores de alto risco—a margem de $5,6\%$ entre BERTweet e ALBERT pode ser inaceitável, justificando o custo adicional.

Uma limitação importante do presente estudo é que não avaliamos técnicas de quantização em tempo de inferência, que poderiam reduzir significativamente o custo de *deployment* dos modelos já treinados, sem impacto substancial na *performance*. Estudos futuros devem investigar esse cenário.

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma análise comparativa empírica entre duas estratégias proeminentes para a otimização de modelos de linguagem *Transformer*: a compressão de modelos e a especialização de domínio. Ao avaliar três arquiteturas representativas—DistilBERT, ALBERT e BERTweet—na tarefa de análise de sentimentos em *tweets*, o estudo buscou quantificar o *trade-off* entre o desempenho preditivo e o custo computacional, preenchendo uma lacuna significativa na literatura onde essas dimensões raramente são avaliadas conjuntamente.

Os resultados confirmaram a hipótese de que a especialização de domínio, representada pelo BERTweet, é a estratégia superior para alcançar uma boa *performance*, liderando em acurácia ($84,47\%$), Macro-F1 ($84,38\%$) e precisão ($85,31\%$). Esse achado reforça a importância do alinhamento entre o *corpus* de treinamento e o domínio de aplicação, especialmente para linguagem informal e ruidosa como a de redes sociais.

Por outro lado, a análise de custo-benefício revelou que modelos compactos, notadamente o DistilBERT, oferecem uma alternativa altamente viável, entregando um desempenho robusto com uma fração significativa dos recursos computacionais (50% dos FLOPs e 45% do tempo de treinamento) de seu análogo especializado. Para organizações operando sob restrições orçamentárias ou ambientais, essa opção representa um equilíbrio pragmático entre capacidade preditiva e sustentabilidade operacional.

O ALBERT, embora seja a arquitetura mais econômica em termos de parâmetros (12 M) e FLOPs ($0,46 \times 10^{15}$), demonstrou uma queda de *performance* mais acentuada ($78,84\%$ de acurácia), evidenciando os limites da compressão agressiva para tarefas complexas de classificação de texto em domínios ruidosos. Este resultado sugere que existe um ponto de diminuição marginal na relação entre redução paramétrica e retenção de *performance*.

A principal contribuição deste estudo é a apresentação de direções empíricas para profissionais enfrentando decisões de *trade-off* entre especialização e eficiência. A escolha entre um modelo compacto ou especializado não deve ser baseada apenas na *performance* máxima, mas sim nas restrições e nos objetivos da aplicação final:

- Para cenários críticos de desempenho: BERTweet é a escolha preferencial quando a precisão é imperativa e recursos computacionais não constituem fator limitante.
- Para cenários de custo-eficiência otimizada: DistilBERT emerge como a opção mais adequada, oferecendo uma redução em requisitos computacionais com degradação mínima de *performance*.

- Para aplicações com restrições severas de recursos: ALBERT pode ser apropriado para cenários específicos de *edge computing*, aceitando a degradação de *performance* em troca de viabilidade de *deployment*.

Reconhecemos algumas limitações do presente trabalho que devem ser endereçadas em futuras investigações. Este estudo não explorou técnicas de quantização pós-treinamento, que podem reduzir significativamente o tamanho e a latência dos modelos sem impacto substancial na *performance*, sendo necessário que futuras pesquisas integrem essas técnicas para obter uma visão mais completa do *landscape* de otimização.

A avaliação foi limitada ao Sentiment140 em inglês, sendo importante expandir para outros *corpora* de redes sociais como Reddit, TikTok, Instagram, além de linguagens não-inglesas e domínios relacionados como análise de emoções, detecção de toxicidade e outras análises cruciais para generalizar os achados.

Investigações futuras devem explorar estratégias de *ensemble* que combinem os pontos fortes de múltiplos modelos ou arquiteturas híbridas que adaptem dinamicamente entre compressão e especialização conforme o contexto. Modelos comprimidos frequentemente apresentam degradação em calibração de incerteza, exigindo pesquisas futuras para avaliar como diferentes estratégias afetam a confiabilidade das predições.

Examinar se aplicar pré-treinamento contínuo em dados de domínio específico aos modelos comprimidos, como DistilBERT e ALBERT, poderia aproximar seu desempenho ao BERTweet enquanto mantém a eficiência também representa uma direção promissora. Finalmente, análises de como os modelos respondem a ruído adicional e perturbações podem fornecer *insights* sobre quando cada estratégia é mais robusta.

Este trabalho estabelece, portanto, um ponto de partida para futuras investigações sistemáticas na interseção de especialização de domínio, compressão de modelos e eficiência computacional em tarefas de PLN prático. Esperamos que os achados e metodologia aqui apresentados informem desenvolvimentos futuros em modelos de linguagem mais sustentáveis e acessíveis para análise de redes sociais em escala.

Referências

Adu-Acheampong, F., Okyere-Dankwa, B., and Yeboah-Akowuah, B. (2021). Transformer models for text-based sentiment analysis: a review of bert-based approaches. *Journal of Smart Systems and Technologies*, 1(1):1–16.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Bhatt, N., Patel, K., Chotai, Y., and Shah, V. (2020). A comparative study of pre-trained language models for task-oriented dialogue system. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 487–492, Suzhou, China. Association for Computational Linguistics.

Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., and Guttag, J. (2020). What is the state of neural network pruning? *Proceedings of Machine Learning and Systems*, 2:129–146.

Cai, H., Gan, C., Wang, T., Zhang, Z., and Han, S. (2020). Once-for-all: Train one network and specialize it for efficient deployment. In *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:1908.09791.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Frankle, J. and Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ganesh, P., Chen, Y., Le, D., Suda, N., Su, H., and M., R. (2021). Compressing BERT: A case study in Deep Model Compression. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7.

Go, A., Bhayani, R., and Huang, L. (2009a). Twitter sentiment analysis. Technical report, Stanford University. CS224N Project Report.

Go, A., Bhayani, R., and Huang, L. (2009b). Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint*, arXiv:1503.02531.

Kavitha, P., Janaki, D., Ayathri, M., Kirubatini, G., and Kaveri (2025). Applications, challenges, and emerging trends of twitter sentiment analysis. *Zenodo*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Nguyen, D. Q., Vu, T., and Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1320–1326.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Patterson, D. A., Gonzalez, J., Le, Q. V., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint*, arXiv:2104.10350.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Ramesh, K., Chavan, A., Pandit, S., and Sitaram, S. (2023). A comparative study on the impact of model compression techniques on fairness in language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada. Association for Computational Linguistics.

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune BERT for text classification? In *Chinese Computational Linguistics*, pages 194–206. Springer International Publishing.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022a). Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022b). Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5998–6008. Curran Associates, Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xie, Y., Aggarwal, K., and Ahmad, A. (2024). Efficient continual pre-training for building domain specific large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.

Zhou, H., Lan, J., Liu, R., and Yosinski, J. (2019). Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.