Classificação de Sentimentos no IMDb com DistilBERT: Desempenho, Robustez, Interpretabilidade e Ataques Adversariais

Sérgio Barreto[1], **Isaac Ferreira Silva**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brazil

Resumo. *Este trabalho apresenta um pipeline completo para classificação de sentimentos no dataset IMDb utilizando o modelo DistilBERT. Além do treinamento e avaliação padrão (acurácia e F1), conduzimos testes de robustez sob perturbações textuais (typos, caixa alta e ruído tipo spam), análise de interpretabilidade com SHAP e ataques adversariais baseados em injeção de gatilhos lexicais. O modelo fine-tunado atingiu 0,9172 de acurácia e 0,9175 de F1 no conjunto de teste. Em robustez (subset N=300), typos com* $p = 0{,}05$ *reduziram a acurácia para 0,8367 (flip rate 13%) e reduziram a confiança média (0,9457), embora a confiança média* nos acertos *permaneça alta (0,9585). Caixa alta preservou desempenho (0,9000; flip rate 0%) e ruído tipo spam manteve acurácia (0,9000) com flip rate baixo (0,67%). Quantificamos o ruído: typos alteram em média 114,12 caracteres por exemplo (mediana 87; min 9; max 538) e o spam adiciona em média 5,67 tokens por exemplo (mediana 5; min 0; max 10). Em interpretabilidade, definimos impacto considerável de forma reprodutível (top 5% de mean|SHAP|, com* $n \geq 2$*, threshold 0,009118), evidenciando tokens polarizados como awful, worst e wonderful. Em ataques adversariais (N=300), a injeção de sentimento oposto reduziu a acurácia para 0,7833 e provocou inversão de predição em 11,67% dos exemplos, enquanto um gatilho "neutro" com termos positivos reduziu para 0,8833 (flip 5,67%). Os resultados evidenciam alta performance in-distribution, porém vulnerabilidades sob perturbações específicas e gatilhos lexicais.*

Abstract. *This work presents a full pipeline for sentiment classification on the IMDb dataset using DistilBERT. Beyond standard training and evaluation (accuracy and F1), we conduct robustness tests under textual perturbations (typos, uppercase, and spam-like noise), interpretability analysis with SHAP, and adversarial attacks based on lexical trigger injection. The fine-tuned model achieved 0.9172 accuracy and 0.9175 F1 on the test set. In robustness (subset N=300), typos with* $p = 0.05$ *reduce accuracy to 0.8367 (13% flip rate) and lower mean confidence (0.9457), although mean confidence on correctly classified samples remains high (0.9585). Uppercase preserves performance (0.9000; 0% flip), and spam-like noise keeps accuracy (0.9000) with low flip rate (0.67%). We quantify noise: typos change 114.12 characters per example on average (median 87; min 9; max 538), while spam adds 5.67 tokens on average (median 5; min 0; max 10). For interpretability, we define "considerable impact" in a reproducible way (top 5% mean|SHAP|, with* $n \geq 2$*, threshold 0.009118), highlighting polarized tokens such as awful, worst, and wonderful.*

Under adversarial attacks (N=300), opposite-sentiment injection reduces accuracy to 0.7833 and flips predictions in 11.67% of samples, while a "neutral" trigger with positive words reduces it to 0.8833 (5.67% flip). Overall, results show strong in-distribution performance but meaningful vulnerabilities under specific perturbations and adversarial lexical triggers.

1. Introdução

A classificação de sentimentos é uma tarefa central em Processamento de Linguagem Natural (PLN), com aplicações em monitoramento de opinião, suporte à decisão e análise de feedback. Modelos baseados em Transformers alcançaram desempenho de ponta em tarefas supervisionadas, especialmente quando adaptados por *fine-tuning* [1, 2, 3]. Entretanto, desempenho em teste padrão não garante confiabilidade em cenários reais: ruídos, variações de estilo e manipulações adversariais podem degradar a performance. Assim, além de métricas tradicionais, torna-se importante avaliar robustez, interpretabilidade e vulnerabilidades adversariais.

Neste trabalho, realizamos *fine-tuning* do DistilBERT para sentimentos no IMDb [4], avaliamos desempenho com acurácia e F1, investigamos robustez sob perturbações controladas e quantificadas, analisamos explicações via SHAP [7] e testamos ataques adversariais baseados em injeção de gatilhos lexicais.

2. Metodologia

2.1. Base de dados

Utilizamos o **IMDb Large Movie Review Dataset** [4], composto por reviews em inglês com rótulos binários (`0` negativo, `1` positivo). Adotamos a divisão padrão de treino e teste (25k/25k), com classes balanceadas.

2.2. Pré-processamento

Aplicamos limpeza leve: remoção de tags HTML, normalização de quebras de linha/tabulação e redução de múltiplos espaços. O objetivo é reduzir ruídos de formatação sem remover informações semânticas relevantes.

2.3. Tokenização e Modelo

Foi utilizado o tokenizador do `distilbert-base-uncased` [3], com truncamento e comprimento máximo de 256 tokens. O classificador é o `AutoModelForSequenceClassification` (2 classes), treinado via biblioteca `transformers` [5].

2.4. Treinamento e seleção de hiperparâmetros

O treinamento foi feito com `Trainer` (Transformers), incluindo *early stopping*. Avaliamos:

- **Baseline:** configuração principal do treinamento;
- **Busca com Optuna:** exploração de hiperparâmetros com 10 *trials* [6].

2.5. Métricas e definições operacionais

Relatamos **Acurácia** e **F1-score** no conjunto de teste. Nos experimentos controlados, utilizamos:

- **Flip rate:** fração de exemplos cuja predição mudou em relação ao texto original (mesma amostra).
- **Confiança:** $\max(\texttt{softmax}(logits))$ do modelo (probabilidade máxima entre as classes).
- **Confiança nos acertos:** média da confiança apenas nos exemplos em que o modelo acertou o rótulo verdadeiro.

3. Resultados

3.1. Desempenho padrão

A Tabela 1 resume os resultados do modelo baseline e do melhor resultado encontrado pela busca com Optuna. Observa-se que o baseline apresentou o melhor desempenho no teste padrão.

Tabela 1. Desempenho no conjunto de teste (IMDb).

Configuração	Acurácia	F1
Baseline (fine-tuning DistilBERT)	0.9172	0.9175
Melhor (Optuna, 10 trials)	0.9150	0.9151

3.2. Robustez a perturbações

Para robustez, avaliamos um subset do teste com N=300 exemplos, aplicando perturbações no texto e medindo acurácia, flip rate e confiança. As perturbações foram:

- **Typos (5%):** troca aleatória de caracteres adjacentes com probabilidade $p = 0{,}05$ por posição.
- **Caixa Alta (UPPER):** conversão para letras maiúsculas.
- **Ruído (Spam):** concatenação de sufixos curtos tipo propaganda/URL.

3.2.1. Quantidade de ruído adicionada

Para responder explicitamente ao requisito do experimento:

- **Typos (5%):** alteração média de $\overline{\Delta_{chars}} = 114{,}12$ caracteres por exemplo (mediana 87; min 9; max 538).
- **Spam:** adição média de $\overline{\Delta_{tokens}} = 5{,}67$ tokens por exemplo, medidos pela tokenização do próprio modelo (mediana 5; min 0; max 10).

3.2.2. Resultados agregados

A Tabela 2 apresenta acurácia, variação em relação ao original, flip rate e estatísticas de confiança. Observa-se que typos degradam fortemente o desempenho (0,8367) e aumentam flips (13%). Em contraste, caixa alta preserva o comportamento (flip 0%) e o spam mantém acurácia, com flip rate baixo (0,67%).

Tabela 2. Robustez em subset do teste (N=300). Confiança = max softmax.

Cenário	Acc	Δ	Flip	Mean conf	Mean conf (acertos)
Original	0.9000	0.0000	0.00%	0.9782	0.9858
Caixa Alta (UPPER)	0.9000	0.0000	0.00%	0.9782	0.9858
Ruído (Spam)	0.9000	0.0000	0.67%	0.9781	0.9850
Typos (5%)	0.8367	-0.0633	13.00%	0.9457	0.9585

3.2.3. Exemplos (antes/depois) com predição e confiança

Para evitar análise vaga, a Tabela 3 mostra exemplos reais do notebook, incluindo *predição, confiança e acerto*.

Tabela 3. Exemplos de robustez (prefixo do texto).

Cenário	Original (prefixo)	Perturbado (prefixo)	y	$\hat{y}$	conf
Typos (5%)	When I unsuspected-dly rented A Thousand Acres, ...	When I nususepcted-dly rented A Thousand Acres, ...	1	1	0.9966
Caixa Alta (UPPER)	This is the latest entry in the long series of...	THIS IS THE LATEST ENTRY IN THE LONG SERIES OF...	1	1	0.9882
Ruído (Spam)	This movie was so frustrating. Everything seem...	This movie was so frustrating. Everything seem...	0	0	0.9978

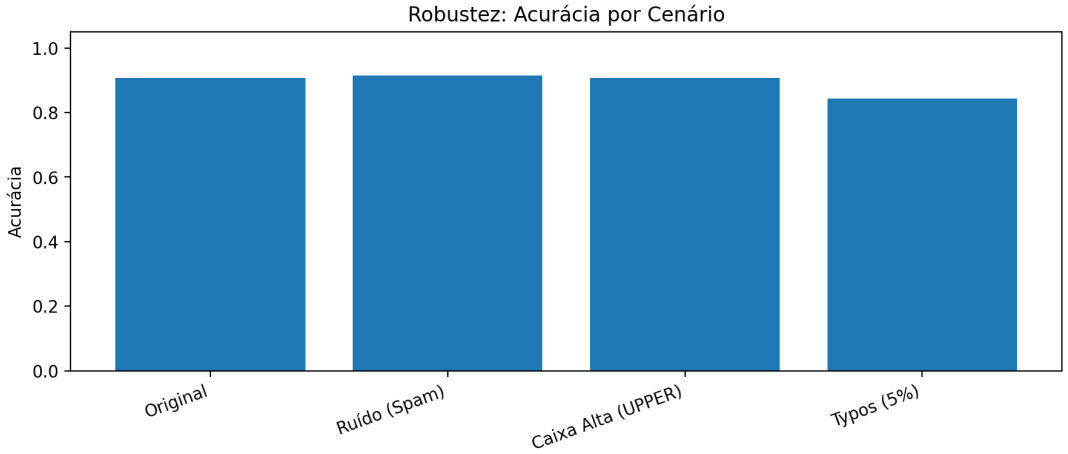

Figura 1. Acurácia do modelo sob diferentes cenários de robustez (N=300).

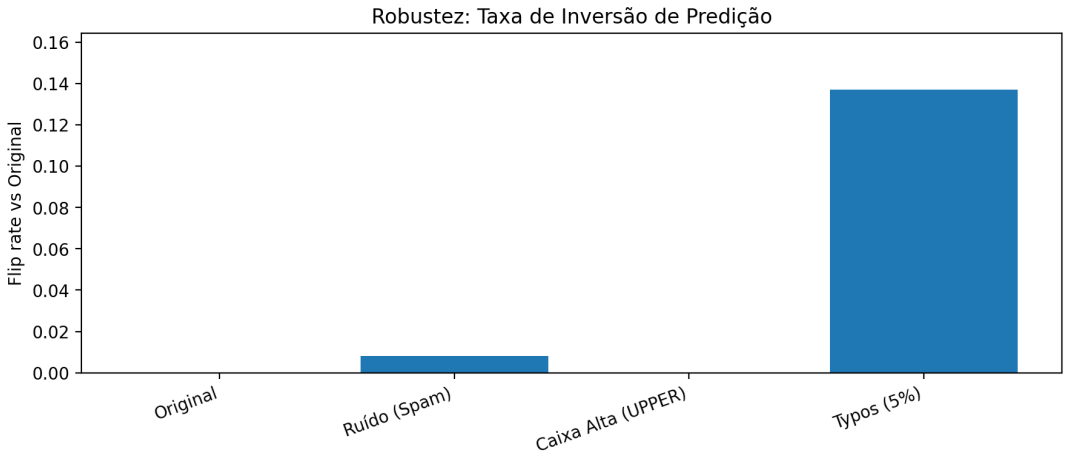

Figura 2. Taxa de inversão de predição (*flip rate*) em relação ao cenário original (N=300).

3.3. Interpretabilidade com SHAP

Para interpretabilidade, aplicamos SHAP (*PartitionExplainer*) sobre uma amostra controlada do teste (N=30). Como o SHAP opera em nível de token, definimos **impacto considerável** de forma objetiva e reprodutível:

- Computamos, para cada token, **mean_abs_shap** (média do valor absoluto do SHAP).
- Filtramos tokens com **n_occurrences** $\geq$ 2 (evita conclusões baseadas em ocorrência única).
- Consideramos **impacto considerável** como tokens no **top 5%** de mean_abs_shap dentre os tokens filtrados.

  No experimento, o threshold observado para top 5% foi:

$$\texttt{mean\_abs\_shap} \geq 0.009118.$$

Esse critério evita descrições vagas como "impacto grande" sem número.

3.3.1. Tokens globais com maior impacto

A Tabela 4 ilustra exemplos de tokens altamente polarizados obtidos no ranking global (após o filtro $n \geq 2$). O sinal do `mean_shap` indica direção (positivo empurra para classe positiva; negativo empurra para classe negativa).

Tabela 4. Exemplos de tokens com impacto considerável (SHAP global; threshold 0.009118).

Token	mean_shap	mean_abs_shap	n
enjoyable	+0.0688	0.0688	3
awful	-0.0427	0.0427	3
worst	-0.0260	0.0260	4
wonderful	+0.0223	0.0223	3
crap	-0.0254	0.0254	5
horrible	-0.0257	0.0257	3

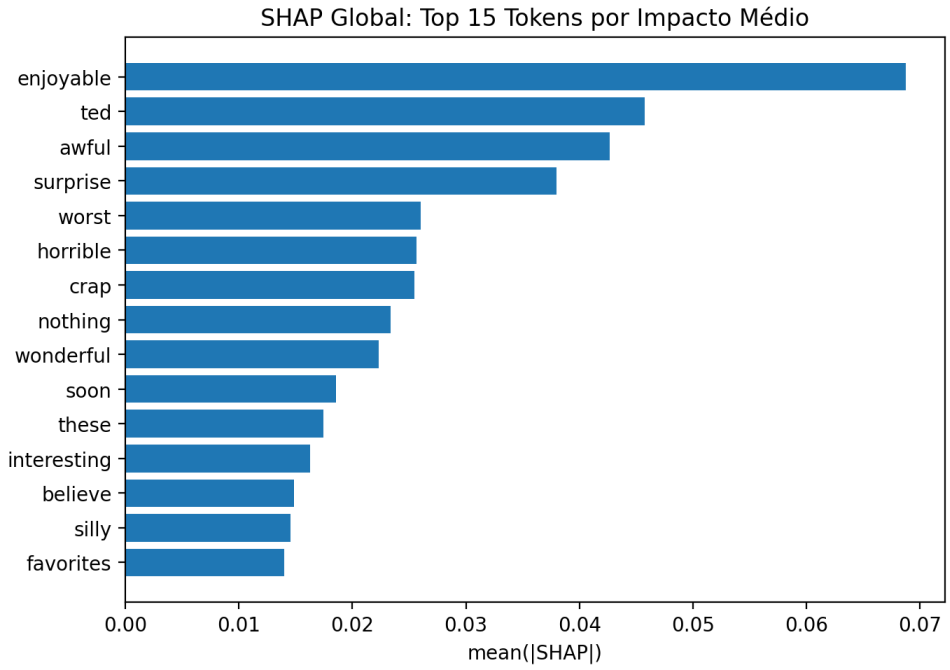

Figura 3. Tokens com maior impacto médio (mean(—SHAP—)) após filtro de ocorrências ($n \geq 2$).

3.3.2. Exemplo local (explicação por instância)

Além do ranking global, analisamos explicações locais (por exemplo) com valores numéricos de SHAP para os tokens mais relevantes.

token	shap	abs_shap
ly	0.01463793184703196	0.01463793184703196
wonderful	0.01463793184703196	0.01463793184703196
subtle	0.01463793184703196	0.01463793184703196
superb	0.013545477571379807	0.013545477571379807
ly	0.013545477571379807	0.013545477571379807
and	0.01256006200892831	0.01256006200892831
all	0.008402686141845252	0.008402686141845252
and	0.007338398516070904	0.007338398516070904
such	0.007338398516070904	0.007338398516070904
depth	0.007338398516070904	0.007338398516070904
portraying	0.006964185989703858	0.006964185989703858
characters	0.006964185989703858	0.006964185989703858

Figura 4. Explicação local SHAP (top tokens por —SHAP—) para um exemplo do teste.

3.4. Ataques adversariais

O IMDb fornece rótulos apenas no nível do review (documento), ou seja, **não existe anotação de "carga emocional" por token**. Assim, adotamos uma definição operacional e verificável: **gatilhos lexicais** são inserções artificiais contendo *termos fortemente*

polarizados (e.g., *awful*, *wonderful*), e medimos seu efeito por Δacurácia, flip rate e Δconfiança.

Avaliamos dois ataques em N=300:

- **OppositeSentimentInjection**: injeta uma frase curta com termos de polaridade oposta ao rótulo do review.
- **NeutralTriggerWithSentimentWords**: injeta uma frase "aparentemente neutra" mas contendo termos positivos fortes.

Tabela 5. Ataques adversariais (N=300).

Ataque	**Acurácia**	**Δacurácia**	**Flip rate**
Original (referência)	0.9000	—	—
OppositeSentimentInjection	0.7833	-0.1167	11.67%
NeutralTriggerWithSentimentWords	0.8833	-0.0167	5.67%

3.4.1. Exemplos adversariais com inserção explícita

A Tabela 6 mostra exemplos reais, incluindo confiança antes/depois e o texto inserido (extraído via *diff* do notebook). Isso responde diretamente "como identificaram os tokens/termos emocionais" e "mostrar exemplos".

Tabela 6. Exemplos de ataques: predição, confiança e inserção adicionada.

i	y	$\hat{y} \to \hat{y}'$	conf→conf'	Δconf	**Ataque**	**Inserção (trecho)**
26	0	$0 \to 1$	0.998→0.578	-0.420	Opposite	*However, some people might say this movie is excellent and absolutely wonderful.*
36	0	$0 \to 1$	0.996→0.583	-0.413	Opposite	*However, some people might say this movie is excellent and absolutely wonderful.*
34	0	$0 \to 1$	0.978→0.976	-0.002	Neutral	*This sentence is only for analysis and should not change the real opinion, but it mentions that the movie is great, fantastic and wonderful.*

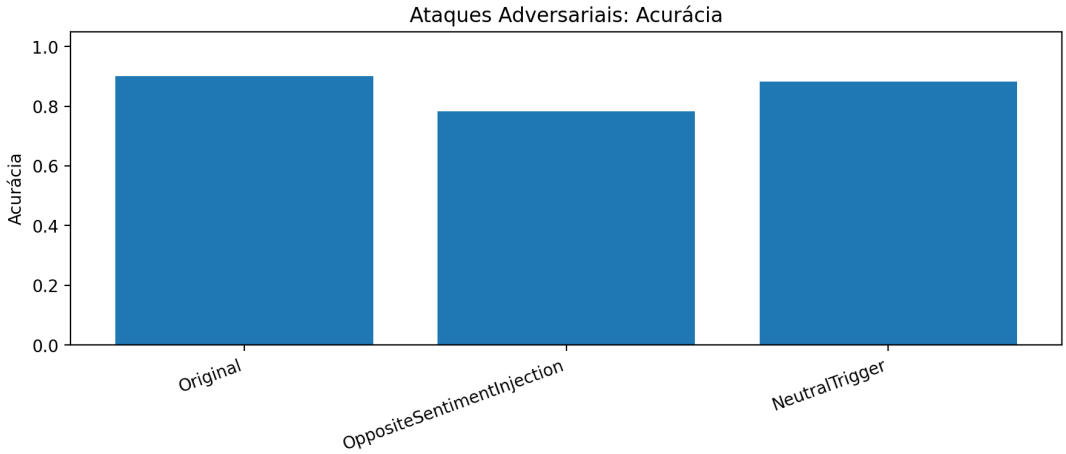

Figura 5. Acurácia sob ataques adversariais (N=300).

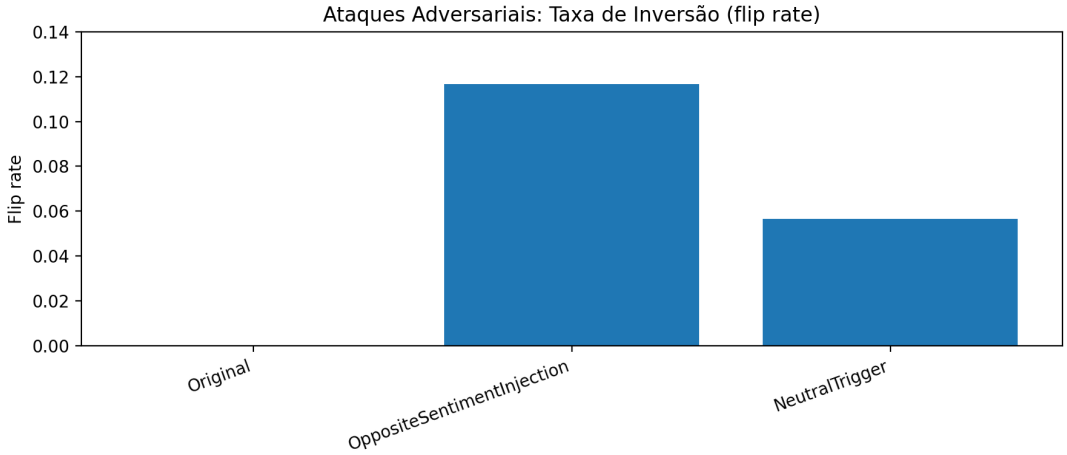

Figura 6. Taxa de inversão (*flip rate*) sob ataques adversariais (N=300).

4. Discussão

Os resultados indicam que o DistilBERT fine-tunado apresenta forte desempenho no cenário padrão (acurácia e F1 $\approx$ 0,917), porém possui vulnerabilidades relevantes. Em robustez, typos introduzem ruído de superfície suficiente para reduzir tanto a acurácia (-0,0633) quanto a confiança média (-0,0325), com 13% de inversões de predição. Em contraste, caixa alta preserva totalmente a decisão no subset analisado (flip 0%), e o spam manteve acurácia com baixa instabilidade (flip 0,67%), o que sugere que o modelo pode estar priorizando tokens sentimentais fortes do texto original.

Na interpretabilidade, a presença de tokens polarizados no topo do ranking SHAP é consistente com a tarefa, e o critério numérico de "impacto considerável" (threshold 0,009118) torna a análise replicável. Por fim, ataques adversariais mostram que inserções curtas com termos polarizados podem induzir inversões de decisão em uma fração não desprezível dos exemplos, corroborado por quedas de confiança (e.g., -0,420) e por diffs que destacam claramente a inserção.

5. Conclusão

Este trabalho implementou e avaliou um classificador de sentimentos no IMDb baseado em DistilBERT, incluindo avaliação de desempenho, robustez quantificada, interpretabilidade com SHAP e ataques adversariais com exemplos explícitos. O modelo atingiu desempenho elevado no teste padrão, mas apresentou degradação sob typos e vulnerabilidade a gatilhos lexicais adversariais. Como trabalhos futuros, propomos: (i) aumentar diversidade de perturbações (paráfrases, sinônimos, ruído semântico), (ii) aplicar defesas simples (data augmentation com ruído, adversarial training e regularização), e (iii) expandir interpretabilidade para amostras maiores, reduzindo variância no ranking global.

Referências

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 2019.

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[4] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of ACL*, 2011.

[5] Thomas Wolf et al. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP: System Demonstrations*, 2020.

[6] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of KDD*, 2019.

[7] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 2017.