CESUPA

Ementa do Primeiro Projeto de Implementação do 2° Bimestre

(AV2): Mecanismos de Atenção e Arquiteturas Transformer

Fine-Tuning de RoBERTa com LoRA para Classificação de Sentimentos no Dataset Rotten Tomatoes: Avaliação e Ablação

BELÉM

2025

CAUÊ MARTINS

GABRIEL ANTONIO

GABRIEL GALEGO

JOÃO VITOR FARIAS

MURILO GUIMARÃES

WALLACE RICARDO

Resumo

Este trabalho investiga o uso de modelos baseados em Transformer para a tarefa de classificação de sentimentos utilizando o dataset Rotten Tomatoes Movie Reviews. A abordagem inclui o treinamento de três modelos distintos: RoBERTa-base, RoBERTa-base adaptado com LoRA (Low-Rank Adaptation) e uma versão ablatada com redução do número de hidden layers e attention heads. Todo o processo segue princípios de reprodutibilidade, rigor metodológico e práticas modernas de ajuste fino, utilizando a API Trainer da Hugging Face. Os resultados indicam que o modelo com LoRA supera a versão base em acurácia e F1-macro, enquanto a versão ablatada apresenta queda significativa, permitindo uma compreensão estruturada sobre como diferentes componentes arquiteturais influenciam o desempenho final.

Palavras-chave— Transformers, RoBERTa, LoRA, Fine-tuning, Rotten Tomatoes.

BELÉM

2025

1. INTRODUÇÃO

A evolução dos modelos de linguagem baseados em Transformers redefiniu o panorama do Processamento de Linguagem Natural, fornecendo arquiteturas mais expressivas e eficientes para tarefas de classificação de texto. Trabalhos pioneiros como BERT e RoBERTa demonstram que mecanismos de atenção são capazes de capturar dependências semânticas complexas com maior precisão do que modelos recorrentes. Contudo, o fine-tuning completo de redes profundas como RoBERTa apresenta custos substanciais de hardware, tempo e consumo de energia, o que motivou o surgimento de técnicas alternativas, como LoRA, que busca adaptar modelos pré-treinados utilizando matrizes de baixa dimensão em pontos estratégicos da arquitetura. O presente relatório investiga de maneira aplicada o comportamento de um Transformer encoder-only baseado no RoBERTa-base, treinado com diferentes estratégias.

O dataset Rotten Tomatoes, contendo avaliações de filmes rotuladas como positivas ou negativas, foi utilizado como conjunto de dados principal, permitindo explorar o desempenho dos modelos em um cenário clássico de classificação. O objetivo do estudo é analisar comparativamente o modelo RoBERTa-base com fine-tuning completo, o modelo RoBERTa adaptado com LoRA, e um modelo ablatado com capacidade reduzida. A experimentação permite avaliar a eficiência computacional, a expressividade da atenção multi-head e o impacto das modificações estruturais no desempenho final. A partir desse problema aplicado, investigou-se como o modelo interage com o conteúdo textual real das avaliações, como essas interações se materializam durante o pré-processamento, tokenização e treinamento, e como diferentes versões do Transformer respondem ao mesmo conjunto de exemplos. O estudo fornece tanto uma perspectiva prática quanto uma análise comparativa detalhada, discutindo pontos fortes e limitações de cada abordagem.

2. FUNDAMENTAÇÃO TEÓRICA

A arquitetura Transformer representa uma ruptura tecnológica ao substituir os mecanismos recorrentes tradicionais pelo conceito de self-attention. O cerne desse mecanismo está na capacidade de cada token de uma sequência textual comparar-se a todos os outros simultaneamente por meio dos vetores Query, Key e Value, produzindo representações contextualizadas a partir da aplicação da função softmax sobre a matriz $QK^T/\sqrt{dk}$. Esse processo, por ser altamente paralelizável, elimina a dependência sequencial de arquiteturas como LSTMs, resultando em modelos mais rápidos e robustos ao capturar relações de longo alcance. O uso de múltiplas cabeças de atenção, característica conhecida como multi-head attention, permite que o modelo explore o espaço semântico por diferentes perspectivas, aprendendo padrões independentes de associação entre tokens.

Dentro dessa estrutura, RoBERTa emerge como uma variante otimizada de BERT, eliminando componentes que se mostraram pouco eficazes, como o next sentence prediction, e aumentando significativamente o volume de dados e o tempo de pré-treinamento. Essa variação resulta em embeddings contextuais mais ricos, tornando RoBERTa-base particularmente adequado para tarefas de classificação. No entanto, o fine-tuning integral de suas camadas internas é custoso em termos computacionais, tornando desejável o uso de técnicas mais eficientes. Nesse sentido, LoRA apresenta uma abordagem inovadora ao inserir matrizes adicionais de baixa dimensão apenas nos módulos de projeção dos vetores Query e Value, mantendo congelados os pesos originais do Transformer. Essa adaptação reduz drasticamente o número de parâmetros treináveis enquanto preserva a capacidade expressiva da rede. A ablação estrutural, presente neste estudo, fornece uma perspectiva complementar, uma vez que reduz intencionalmente o número de camadas e cabeças de atenção no Transformer para investigar o quanto essas características influenciam o desempenho e a generalização.

3. METODOLOGIA

O desenvolvimento metodológico deste trabalho articula, de forma integrada, a preparação dos dados, a construção dos modelos, o treinamento e a avaliação, assegurando que o pipeline seja consistente com boas práticas científicas de reprodutibilidade e rigor experimental. O dataset Rotten Tomatoes retorna automaticamente dividido em conjuntos de treino, validação e teste, contendo avaliações textuais rotuladas como positivas ou negativas. Após o carregamento, cada instância passa por tokenização realizada pelo AutoTokenizer da HuggingFace, configurado para truncamento e padding dinâmico, de forma a garantir eficiência durante o treinamento. A tokenização transforma sequências textuais em vetores inteiros que representam unidades sublexicais do modelo RoBERTa, preservando relações morfológicas e minimizando o problema de palavras desconhecidas. Esse processo é complementado pelo DataCollatorWithPadding, que otimiza a construção dos batches para a GPU.

O modelo RoBERTa-base serve como ponto de partida para o fine-tuning tradicional, sendo carregado com pesos pré-treinados e adaptado por meio de uma nova camada classificadora. Contudo, o foco metodológico repousa na aplicação de LoRA, configurada com rank 8, alpha 32 e dropout 0.1, atuando exclusivamente nas projeções Query e Value das camadas de atenção. A incorporação de LoRA é realizada por meio da biblioteca PEFT, que identifica automaticamente os módulos-alvo e injeta parâmetros adicionais sem modificar ou duplicar a estrutura base do Transformer. A partir dessa modificação, o modelo é treinado com apenas uma fração dos parâmetros ajustáveis, permitindo treinos mais estáveis e menos custosos. Em paralelo, a ablação arquitetural é implementada alterando a configuração interna do Transformer, reduzindo o número total de hidden layers e attention heads, fato que produz um modelo supervisionado com menor capacidade expressiva. Essas três variantes — base, LoRA e ablatada são treinadas utilizando a API Trainer, que coordena o processo de forward, backward, otimização e early stopping. A escolha dos hiperparâmetros, como taxa de aprendizado de 2e-5, peso de decaimento 0.01 e até 10 épocas de treino, foi guiada por práticas consolidadas no treinamento de modelos encoder-only para classificação. Como forma de garantir reprodutibilidade, foram fixadas seeds para todos os módulos relevantes, incluindo NumPy, PyTorch e CUDA.

4. RESULTADOS

Os resultados obtidos refletem diretamente a forma como cada variante do modelo interage com os textos do Rotten Tomatoes e como interpreta sinais linguísticos associados à polaridade. O RoBERTa-base obteve desempenho consistente, com acurácia e F1-macro próximas a 0.92, produzindo classificações estáveis mesmo em frases curtas e subjetivas. Durante o processo de avaliação, observou-se que o modelo base lida bem com críticas positivas explícitas, mas comete erros ocasionais em resenhas cuja opinião é mais ambígua ou contém construções irônicas — comuns no estilo crítico de cinema.

O modelo com LoRA superou o desempenho do modelo base, alcançando métricas aproximadamente na faixa de 0.94 para acurácia e F1-macro. Essa melhoria se refletiu também na análise qualitativa: nas críticas onde o RoBERTa-base apresentava incertezas, o modelo com LoRA frequentemente fornecia classificações mais coerentes com o rótulo real. O fato de o modelo com LoRA treinar um subconjunto reduzido de parâmetros parece ter contribuído para maior estabilidade e menor tendência ao sobreajuste, permitindo interpretações mais sensíveis quanto à tonalidade emocional das críticas dentro do dataset.

Em contraste, o modelo ablatado apresentou desempenho inferior, com métricas em torno de 0.84 a 0.85. Embora ainda eficaz no básico, ele demonstrou dificuldades evidentes em capturar nuances textuais. Isso ficou claro na análise comparativa gerada pelo código, onde críticas com tons mistos, sarcasmo ou descrições narrativas mais profundas frequentemente eram classificadas incorretamente pelo modelo ablatado. Essa diferença destaca a importância da profundidade e da largura da atenção multi-head para compreender textos subjetivos como os do Rotten Tomatoes. As três séries de resultados foram validadas por meio da função trainer.predict(), seguida de cálculos de accuracy_score e f1_score, garantindo que a comparação entre os modelos fosse conduzida de maneira justa e consistente com o pipeline implementado.

5. DISCUSSÃO

A comparação entre os modelos evidencia como a estrutura arquitetural e a estratégia de treinamento influenciam diretamente a capacidade do sistema de interpretar sentimentos em críticas cinematográficas. O RoBERTa-base mostrou-se altamente competente, mas sua abordagem de fine-tuning completo pode levar a ajustes excessivos em partes da rede que já capturam relações linguísticas relevantes. Por outro lado, o modelo com LoRA demonstrou que adaptar seletivamente apenas determinados módulos — especialmente aqueles associados à projeção de atenção — pode resultar em maior generalização. Uma explicação plausível é que LoRA atua como uma forma de regularização arquitetural, preservando a riqueza do pré-treinamento original e guiando o modelo a incorporar informações específicas do Rotten Tomatoes sem sobrescrever conhecimento linguístico previamente adquirido.

O comportamento do modelo ablatado reforça a hipótese de que o número de camadas e cabeças de atenção não é apenas uma escolha arquitetural arbitrária, mas um elemento chave na capacidade do Transformer de lidar com textos subjetivos e altamente contextuais. Críticas de cinema frequentemente dependem de pistas semânticas dispersas ao longo da frase, jogos de linguagem, metáforas e tonalidades implícitas. Reduzir camadas ou heads limita o número de subespaços nos quais essas relações podem ser modeladas, prejudicando a interpretação global que o modelo forma da crítica.

Outro ponto relevante observado durante a análise foi a diferença qualitativa nas previsões. O código gera tabelas comparativas que permitiram observar que o modelo LoRA superou tanto o modelo base quanto o ablatado em críticas ambíguas. Isso sugere que a adaptação leve não apenas melhora a generalização estatística, mas também fortalece a capacidade interpretativa do modelo no domínio específico do Rotten Tomatoes. Considerando o caráter subjetivo das avaliações e a prevalência de estilos irônicos e informais, essa diferença é particularmente relevante.

Assim, os resultados apoiam a interpretação de que LoRA é uma alternativa eficiente não apenas por reduzir custo computacional, mas também por preservar características fundamentais do pré-treinamento, evitando sobreajuste e permitindo um aprendizado mais direcionado ao domínio. Já a ablação mostrou claramente como a redução de capacidade afeta a sensibilidade do modelo a nuances textuais específicas, enfatizando a importância da arquitetura completa em tarefas semânticas profundas.

6. CONCLUSÃO

Os experimentos realizados demonstram a efetividade dos modelos Transformer na tarefa de classificação de sentimentos e, mais especificamente, evidenciam as vantagens de técnicas modernas de ajuste fino, como LoRA. O RoBERTa-base estabeleceu uma baseline robusta, mas o modelo com LoRA superou-o consistentemente, oferecendo melhor desempenho com menor custo computacional. A ablação arquitetural confirmou que reduções estruturais tendem a prejudicar a capacidade do modelo, embora mantenham desempenho relativamente alto. Esses achados reforçam a relevância dos mecanismos de atenção e das técnicas de adaptação leve para aplicações práticas de NLP. Como continuação deste trabalho, seria interessante examinar diferentes valores de rank para LoRA, realizar buscas automáticas de hiperparâmetros e explorar variantes mais recentes da família RoBERTa e BERT, como DeBERTa e RoBERTa-large.