Relatório Técnico – Classificação Automática de Artigos do arXiv (AI vs ML)

Baseado no colab:

1. Introdução

Este projeto aborda o desenvolvimento e a comparação de três abordagens modernas de classificação automática de textos científicos, focadas nas categorias **cs.AI** (Artificial Intelligence) e **cs.LG** (Machine Learning) do repositório arXiv. O objetivo é investigar a eficiência de métodos contemporâneos baseados em modelos de linguagem avançados, considerando soluções que não exigem dados rotulados, soluções supervisionadas e uma abordagem híbrida baseada em recuperação semântica. A proposta engloba desde o download e preparação dos dados até a avaliação comparativa de desempenho dos métodos utilizados, buscando identificar vantagens, limitações e aplicações práticas para cada técnica.

2. Referencial Teórico

A classificação de textos com modelos de linguagem envolve diferentes níveis de especialização. O primeiro conceito relevante são os modelos **zero-shot**, representados aqui pelo BART-MNLI, que permitem classificar textos sem qualquer treinamento prévio no domínio específico. Eles se baseiam em inferência textual (NLI) e possibilitam prototipagem rápida, embora apresentem dificuldades quando as classes são semanticamente próximas, como ocorre entre AI e ML.

O segundo conceito refere-se ao **fine-tuning**, no qual um modelo pré-treinado é ajustado para uma tarefa específica usando dados rotulados. O modelo SciBERT, especializado em linguagem científica, é especialmente adequado para esse tipo de aplicação, por ter sido treinado sobre grandes volumes de textos acadêmicos. O fine-tuning tende a oferecer o melhor desempenho absoluto, ao custo de maior processamento e necessidade de dataset anotado.

O terceiro conceito é o paradigma **RAG (Retrieval-Augmented Generation/Classification)**. Essa abordagem combina busca semântica com inferência, permitindo que modelos baseados em zero-shot tenham acesso a informações adicionais antes de realizar a classificação. No projeto, isso é implementado por meio de embeddings gerados pelo modelo all-mpnet-base-v2 e de um índice vetorial FAISS, que permite recuperar artigos similares para enriquecer o contexto antes da classificação. Trata-se de uma solução intermediária, útil quando há pouco ou nenhum dado rotulado disponível.

3. Metodologia

O projeto inicia com a coleta automática de artigos do arXiv, utilizando a API oficial. Foram selecionados **500 artigos de cs.AI** e **500 artigos de cs.LG**, totalizando 1000 documentos. De cada artigo foram extraídos título, resumo, categorias, data de publicação e o identificador único. Após a coleta, os dados foram consolidados em um único dataset, com rotulagem binária “AI” ou “ML”.

A preparação do conjunto de dados consistiu em unir título e resumo em um único campo de texto e eliminar inconsistências. Após essa etapa, o dataset foi dividido de forma estratificada em 80% para treino e 20% para teste.

A primeira abordagem aplicada foi o **zero-shot classification** com o modelo BART-MNLI. Para reduzir o custo computacional, foi utilizada uma amostra de 300 textos do conjunto de teste. O classificador recebeu como entradas os textos originais e as duas classes possíveis ("AI" e "ML"), retornando a categoria mais provável para cada documento.

A segunda abordagem utilizou **fine-tuning** do modelo SciBERT. Os textos foram tokenizados com limite de 256 tokens, convertidos para o formato esperado pelo HuggingFace Trainer e submetidos a três épocas de treinamento. Métricas como acurácia, F1 macro, precisão e recall foram calculadas para medir o desempenho do modelo.

A terceira abordagem adotada no projeto foi a **classificação baseada em RAG**. Inicialmente, foram gerados embeddings para todos os textos do corpus utilizando o modelo all-mpnet-base-v2. Em seguida, esses vetores foram indexados utilizando FAISS (Facebook AI Similarity Search), possibilitando a recuperação rápida dos artigos mais similares. Para cada texto do conjunto de teste, foram consultados os cinco vizinhos mais próximos, e o conteúdo desses artigos foi usado para compor um bloco de contexto adicionado ao texto original. Em seguida, esse conjunto enriquecido foi submetido ao mesmo modelo de zero-shot, agora com acesso a informações adicionais provenientes de artigos semelhantes.

4. Resultados

Os resultados obtidos mostram diferenças marcantes entre as abordagens. O **zero-shot** apresentou desempenho satisfatório considerando que não utilizou nenhum dado de treino; porém, enfrentou limitações devido à proximidade semântica entre AI e ML, gerando confusões comuns. Por outro lado, o **fine-tuning do SciBERT** apresentou o melhor desempenho geral. Como o modelo já é especializado em textos científicos e foi ajustado especificamente para distinguir AI de ML, seus resultados superaram as demais técnicas tanto em acurácia quanto em F1 macro.

A abordagem **RAG** obteve desempenho intermediário. Embora não tenha alcançado o SciBERT, superou o zero-shot puro em vários aspectos, demonstrando que a inclusão de contexto externo via recuperação semântica pode melhorar significativamente a classificação quando não se deseja ou não se pode realizar fine-tuning. Em termos qualitativos, o RAG mostrou-se mais interpretável, pois as decisões estavam fundamentadas em artigos similares recuperados do corpus.

Os resultados comparativos sintetizaram o seguinte:

- O zero-shot é rápido e versátil, porém limitado.
- O fine-tuning oferece o desempenho superior, mas requer dados anotados e maior custo computacional.
- O RAG representa um meio-termo eficiente, oferecendo ganhos relevantes sem a necessidade de treinamento adicional.

5. Conclusão

O experimento demonstrou que modelos de linguagem podem ser aplicados de maneira flexível em tarefas de classificação científica, adaptando-se a diferentes restrições de dados e recursos. A solução de zero-shot funciona bem para protótipos rápidos ou ambientes nos quais não há dados rotulados disponíveis. O fine-tuning, por sua vez, é a escolha ideal quando o objetivo é alcançar a melhor performance possível no domínio científico. Já o método RAG se destaca por combinar eficiência e pragmatismo, tornando-se útil em cenários nos quais é necessário contextualizar decisões sem recorrer a treinamento supervisionado.

Do ponto de vista técnico e acadêmico, o projeto evidencia como estratégias modernas de IA generativa e recuperação semântica podem ser combinadas para lidar com problemas de classificação em domínios complexos, como o de artigos científicos. Além disso, demonstra que a escolha da abordagem ideal depende não apenas da busca por performance máxima, mas também das restrições operacionais e dos objetivos do projeto.

6. Mini-Plano de Negócios – Aplicação Comercial da Solução

A solução apresentada possui potencial para se transformar em uma plataforma voltada para instituições acadêmicas, centros de pesquisa e empresas que lidam com grande volume de relatórios técnicos ou documentação científica. O sistema poderia oferecer classificação automática, busca semântica de alta precisão e análise contextualizada dos documentos, usando uma combinação dos métodos testados neste projeto.

A proposta de valor estaria centrada em acelerar a organização e exploração de acervos científicos, garantindo que pesquisadores encontrem rapidamente materiais relevantes e compreendam padrões emergentes nas publicações. O sistema poderia ser oferecido como um serviço baseado em assinatura, com níveis de acesso diferenciados. A versão básica poderia incluir classificação zero-shot e busca simples; a versão intermediária poderia incorporar RAG e buscas semânticas avançadas; enquanto a versão corporativa disponibilizaria fine-tuning personalizado, relatórios analíticos integrados e APIs para integração com sistemas legados.

O desenvolvimento de um MVP seria direto: bastaria permitir upload de PDFs, geração instantânea de embeddings, classificação automática do conteúdo e disponibilização de uma interface para consulta e exploração semântica. Esse MVP poderia ser entregue em poucas semanas e testado inicialmente com universidades ou empresas de tecnologia com alto volume de publicações internas. A escalabilidade do projeto permitiria expandi-lo para outras áreas, como medicina, direito, engenharia e consultorias técnicas, multiplicando as aplicações comerciais.