Projeto Semestral - Ciência de Dados
Classificação de artigos científicos do site Arxiv

Integrantes
Nome: Carolina Perez **RA:** 20.00968-1
Nome: Guilherme Lins Banzato **RA:** 20.01561-5

Github: https://github.com/GuiBanzato/ProjetoCD

Resumo

Este projeto apresenta uma abordagem de aprendizado de máquina para prever a área de conhecimento de artigos científicos, com base em dados do repositório arXiv. Utilizando uma base de 1200 instâncias, o modelo tem como objetivo categorizar artigos em diferentes classes (áreas do conhecimento) a partir do conteúdo textual dos resumos. As técnicas aplicadas incluem embeddings de texto, seleção de características e algoritmos de classificação. Foram utilizados modelos baseados em árvores de decisão para o desenvolvimento do trabalho, especificamente *Random Forest Classifier* e *Decision Tree Classifier*. Uma interface interativa foi desenvolvida em Streamlit, permitindo ao usuário fazer previsões de classes para novos artigos com base no resumo, oferecendo uma aplicação prática para pesquisadores e todos aqueles interessados em organizar e buscar informações relevantes no acervo do arXiv.

Introdução e Contextualização

Com o crescimento exponencial de publicações científicas, tornou-se um desafio organizar e classificar automaticamente artigos em bases de dados acadêmicas. A plataforma arXiv, um dos maiores repositórios de preprints científicos, reúne artigos de diversas áreas, como física, matemática, ciência da computação, entre outras. A classificação correta desses artigos em categorias apropriadas facilita o processo de recuperação de informação e a navegação pelo vasto acervo disponível.

O objetivo deste projeto é desenvolver uma solução de aprendizado de máquina que permita a classificação de artigos científicos por meio de técnicas de processamento de linguagem natural (NLP) e aprendizado supervisionado. Utilizando o conteúdo textual dos resumos dos artigos, foi criado um modelo preditivo capaz de classificar os artigos pelas seguintes categorias: Inteligência Artificial (IA), Cryptography and Security (CR) e Machine Learning (LG). Esta abordagem permite categorizar novos artigos com base em seus resumos, ajudando a encontrar publicações relevantes de forma mais rápida e eficiente.

Metodologia

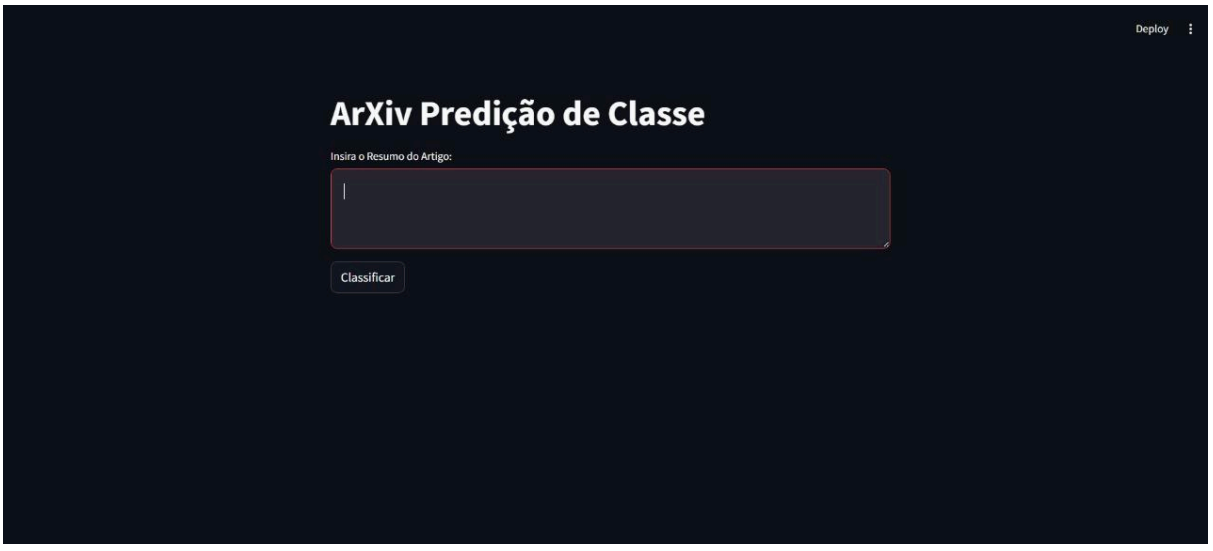
O desenvolvimento do projeto foi dividido nas seguintes etapas:

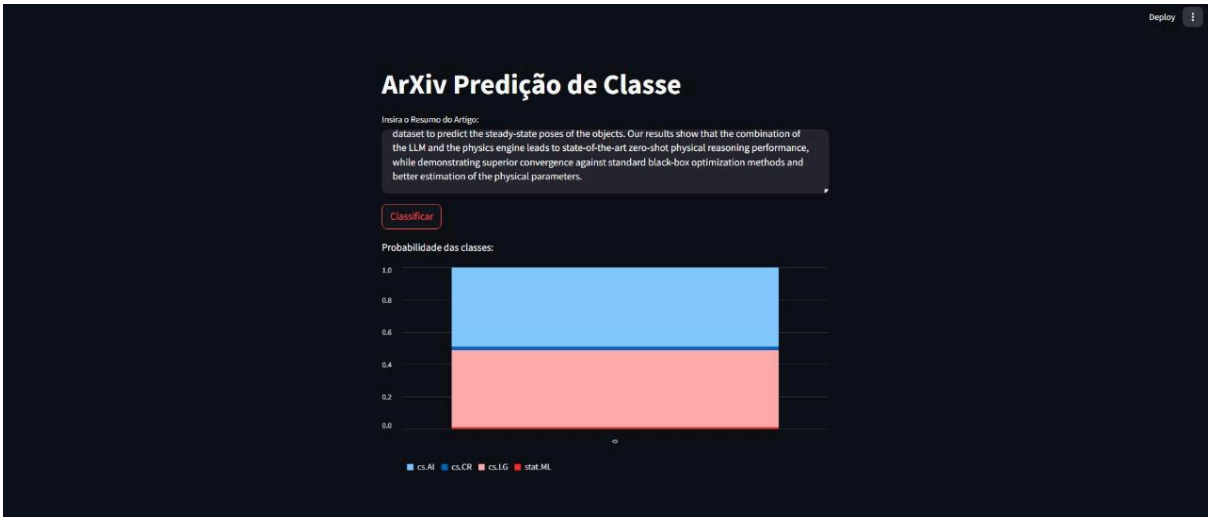
Coleta e Pré-processamento dos Dados: A base de dados utilizada conta com 1200 instâncias e cada dado de entrada inclui o resumo do artigo. O pré-processamento dos textos foi realizado com técnicas de NLP, como tokenização, remoção de stop words e stemming, para melhorar a qualidade dos dados de entrada.

Representação de Dados com Embeddings: Os resumos foram convertidos em embeddings utilizando a técnica *TF-IDF* (Universal Sentence Encoder do TensorFlow), que permite transformar o texto em vetores numéricos que capturam a importância de cada termo no contexto dos artigos científicos. Este vetor numérico serve como entrada para os modelos de classificação.

Desenvolvimento dos Modelos: Foram testados dois algoritmos baseados em árvores de decisão para a tarefa de classificação: *Random Forest Classifier* e *Decision Tree Classifier*. O *Random Forest Classifier* é uma técnica baseada em um conjunto de múltiplas árvores de decisão, que melhora a precisão e reduz a criação de árvores supercomplexas que não generalizam bem os dados (overfitting) ao agregar as previsões das diversas árvores. O *Decision Tree Classifier*, por outro lado, utiliza uma única árvore de decisão para realizar a classificação, o que pode causar o overfitting com mais facilidade caso não haja mecanismos como poda, definição do número mínimo de amostras necessárias em um nó folha ou definição da profundidade máxima da árvore para controlar esse problema. Ambos os algoritmos foram treinados para prever a categoria de um artigo com base nos embeddings dos resumos.

Desenvolvimento da Interface de Usuário: A interface foi implementada com o framework Streamlit, permitindo ao usuário inserir as informações relevantes no método de entrada (resumo) e visualizar as probabilidades de classificação já citadas anteriormente em tempo real.





Resultados e Discussão

O modelo de classificação com *Random Forest Classifier* apresentou uma acurácia média de 80%, enquanto o modelo com *Decision Tree Classifier* obteve acurácia média de 70%, indicando que a combinação de múltiplas árvores (*Random Forest Classifier)* permite uma maior capacidade de generalização. Essa diferença ocorre pois o *Random Forest Classifier*, como já explicado de forma breve anteriormente, combina várias árvores de decisão. Dessa forma, o mesmo usa amostras aleatórias e combina os resultados de diversas árvores, reduzindo o overfitting e resultando em uma melhor acurácia em dados de teste. Portanto, para dados mais complexos, este tipo de modelo tem um desempenho superior ao seu comparativo, neste caso.

A interface em Streamlit facilitou a interação com o sistema, permitindo que os usuários visualizem as probabilidades de classificação para cada categoria por meio de um gráfico de barras empilhadas, esse tipo de gráfico foi escolhido para mostrar a distribuição percentual das três categorias dentro de uma única barra. Essa representação intuitiva das previsões possibilita uma análise rápida das áreas mais prováveis para novos artigos, contribuindo para a descoberta e organização de conhecimento científico.

Conclusão

Este projeto desenvolveu uma ferramenta eficiente e interativa para a classificação automática de artigos científicos em três categorias do arXiv. A aplicação de técnicas de processamento de linguagem natural, juntamente com modelos baseados em árvores de decisão, permitiu obter um sistema de classificação robusto e de fácil utilização, com potencial para auxiliar pesquisadores na organização e busca de informação relevante.

Futuros aprimoramentos podem incluir o aumento da base de dados, adicionar mais categorias de classificação, testar novos modelos e comparar suas

acurácias com os já testados, e por fim, a incorporação de novas variáveis, como citações e autorias, que podem enriquecer a análise e melhorar a precisão dos modelos. Além disso, o uso de embeddings mais sofisticados, como BERT (Bidirectional Encoder Representations), poderia aprimorar ainda mais a capacidade do modelo de captar nuances textuais, uma vez que este foi projetado para pré-treinar representações bidirecionais profundas a partir de texto não rotulado, condicionando conjuntamente o contexto esquerdo e direito em todas as camadas. Como resultado, seu modelo pré-treinado pode ser ajustado com apenas uma camada de saída adicional, criando modelos de última geração para uma ampla gama de tarefas.

Referências

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *arXiv preprint arXiv:1810.04805*. Disponível em: https://arxiv.org/abs/1810.04805 . Acessado em: 13 nov. 2024

arXiv.org. (2024). "arXiv: An Open Access Archive for Research". Disponível em: https://arxiv.org/ . Acessado em: 13 nov. 2014

Ramos, J. (2003). "Using TF-IDF to Determine Word Relevance in Document Queries". *Proceedings of the First Instructional Conference on Machine Learning*.

Breiman, L. (2001). "Random Forests". *Machine Learning*, 45(1), 5–32. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324 . Acessado em: 13 nov. 2024

SCIKIT-LEARN. *Tree-based models*. Scikit-learn: Machine Learning in Python, 2024. Disponível em: https://scikit-learn.org/1.5/modules/tree.html. Acessado em: 13 nov. 2024.