IX Encontro de Iniciação Científica e Tecnológica
IX EnICT
ISSN: 2526-6772
IFSP – *Campus* Araraquara
6 de dezembro de 2025

EM BUSCA DO USO CONSTRUTIVO DA IA EM AMBIENTES EDUCACIONAIS: UM SISTEMA DE DETECÇÃO DE RESPOSTAS GERADAS POR LLMs

RAUAN CARACCIOLO[1], MARCELO CRISCUOLO[2]

[1] Discente do curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de São Paulo (IFSP) no Campus Araraquara. E-mail: rauan.caracciolo@aluno.ifsp.edu.br

[2] Docente do Instituto Federal de São Paulo (IFSP) no Campus Araraquara. E-mail: criscuolo@ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.00.00-7 – Ciência da Computação / Inteligência Artificial

RESUMO: Métodos de detecção automática de textos gerados por LLMs (Large Language Models) podem inibir fraudes e assim contribuir para garantir a integridade do ambiente acadêmico. O propósito deste trabalho é desenvolver um modelo de IA que seja capaz de identificar textos gerados por LLMs com alta precisão e F1-Score utilizando uma abordagem de *black-box detection*. Os processos centrais incluem o treinamento e comparação de dois modelos de aprendizado de máquina, um SVM clássico e uma rede neural multicamadas, suas avaliações e comparação.

PALAVRAS-CHAVE: detecção de textos gerados por IA; LLMs; aprendizado de máquina; integridade acadêmica

INTRODUÇÃO

O uso de Inteligência Artificial (IA), principalmente do tipo de *Large Language Models* (LLM) traz diversos benefícios a seus usuários, como realizar com rapidez pesquisas e análises de dados. Porém, também levanta a possibilidade de fraude em ambientes educacionais e acadêmicos. Métodos de detecção de textos gerados por IAs podem ajudar a resolver esse problema. É necessário, contudo, que esses métodos acompanhem a evolução das próprias IAs (ODRI et al, 2023).

Existem duas técnicas principais para detecção de textos gerados por LLMs: *black-box detection* e *white-box detection*. A primeira é baseada unicamente nos textos gerados pelos LLMs. Esse modelo de detecção se baseia em coletar dados como textos gerados por IAs e textos gerados por humanos, treinar um modelo de ML com eles e utilizá-lo para classificar o texto de *input*. Já o segundo método, o *white-box*, requer acesso completo à aplicação de LLM, pois nesse modo é necessário que os desenvolvedores da IA apliquem uma "marca da água" no texto que faça o detector identificar que o texto foi gerado por aquela IA (TANG et al, 2024).

Neste trabalho, será adotado a abordagem de *black-box detection*, visto que não dispomos de acesso total aos modelos de IA. Inicialmente, será criado um *dataset* contendo textos retirados de obras de domínio público de diversas áreas, como literatura, trabalhos científicos, matérias didáticas, entre outros. Em conjunto com textos gerados por LLMs com temas semelhantes. Após isso, será treinado um modelo SVM clássico e uma rede neural artificial com esses dados e seus resultados alcançados por esses classificadores serão comparados.

FUNDAMENTAÇÃO TEÓRICA

Modelos de Linguagem de Larga Escala (*Large Language Models* – LLMs) são modelos de linguagem generativos. Eles recebem um chamado *prompt* e retornam uma saída para o usuário em forma de texto. O método utilizado para gerar a saída é uma distribuição de probabilidades que o modelo aprendeu e aperfeiçoou em seu treinamento. Ou seja, o LLM apenas acopla palavras que ele reconhece como semelhantes em seu algoritmo treinado (DOUGLAS, 2023). Essa capacidade de mimetismo que impulsiona a utilidade dos LLMs também gera o risco de uso indevido no ambiente acadêmico.

Métodos de detecção mais conhecidos e discutidos são os *black-box detection* e *white-box detection.* Eles se diferem principalmente pelo acesso ao próprio modelo de LLM. Enquanto os métodos *black-box* são aplicados atualmente por meios externos à empresa, pois apenas utilizam dados gerados pelos modelos para treinar outro modelo de classificação, os os métodos *white-box* são um conceito mais difícil de se implementar, pois necessitam que os desenvolvedores por trás do LLM criem um método de detecção interno, como uma "marca da água", que só seria reconhecida por esse algoritmo de detecção (TANG, 2024). O desempenho dos métodos *white-box* é naturalmente superior, uma vez que se apoia na identificação das marcações deixadas pelo próprio modelo generativo. Os métodos *black-box*, por outro lado, dependem totalmente de características linguísticas, extraídas do próprio texto. Essa diferença de funcionamento faz com que os métodos *black-box* estejam mais sujeitos à ocorrência de falsos positivos (classificar textos humanos como gerados por IA) (DALALAH et al, 2023).

Classificação de texto é uma das primeiras áreas criadas em aprendizado de máquina e no processamento de linguagem natural, e tem apresentado bons resultados desde sua criação. As técnicas empregadas para classificação variam de modelos lineares, como SVMs, a modelos conexionistas, como redes neurais multicamadas tradicionais (MLPs) e modelos de aprendizado profundo. A classificação é uma tarefa básica no processamento de linguagem natural. Ela extrai *features* importantes do texto lido e as utiliza para encontrar padrões e aprender a classificá-los. Servindo para diversas aplicações, como prever se uma *review* de produto é positiva ou negativa, prever os sentimentos de uma pessoa pelo modo que ela escreve (LI et al, 2022), e mais especificamente neste trabalho, a detecção de textos gerados por LLMs.

METODOLOGIA

O *dataset* foi criado com base em textos de diversos domínios que estão em domínio público para a classificação de texto gerado por seres humanos. Será desenvolvido um classificador binário com o objetivo de rotular os exemplos de entrada como: Classe 0 (Real), em conjunto com textos gerados por LLMs com propostas semelhantes aos dos humanos para a classificação de texto gerado por IA, Classe 1 (Artificial).

O tratamento dos dados será feito inicialmente deixando todos os caracteres do texto em minúsculas e retirando pontuações. Após, serão aplicados métodos de *word embedding* para extrair features utilizáveis dos textos (KULKARNI et al, 2021). Por fim, o dataset será separado em treinamento e teste, com porcentagem do total de 80% e 20%, respectivamente.

Incorporações de palavras (*word embeddings)* são modelos de representação textual baseados em contexto e redes neurais. Elas representam palavras como vetores densos que capturam relações semânticas entre palavras. Se diferencia do método TF-IDF pois esse se baseia em contagem de palavras e estatística, enquanto o *word embedding* aprende com o contexto.

Na etapa de treinamento, será inicializado um modelo de SVM utilizando a biblioteca de Python *Scikit-learn*, em conjunto com um modelo de rede neural utilizando a biblioteca de Python *Pytorch*. Os modelos serão treinados e avaliados sobre os dados do nosso *dataset.* Para avaliação, usaremos o método de validação cruzada (*k-fold cross validation*).

Na etapa de avaliação, o desempenho dos modelos será comparado utilizando métricas tradicionais de avaliação de modelos de classificação, como Acurácia, Precisão, Recall e F1-Score.

RESULTADOS

A versão inicial do conjunto de dados foi moldada utilizando 6 obras de domínio público, que representam diversas áreas da literatura, como romances, textos acadêmicos, dissertações de filosofia, para a representar a Classe 0, em conjunto, foram utilizados 10 *prompts* de geração de variados, aplicados em 3 modelos de LLMs distintos para representar a Classe 1. Após a coleta dos dados, seus textos foram fragmentados em partes entre 300 e 500 caracteres. Totalizando quinze mil entradas no *dataset*, com um balanceamento de classes relativamente equilibrado.

CONCLUSÃO

O presente trabalho propõe a criação de um modelo de detecção de textos gerados por LLMs, que visa inibir a utilização desses modelos para fraudar trabalhos escolares e acadêmicos, assim incentivando a utilização consciente dessa ferramenta. O trabalho se encontra na fase de testes preliminares do modelo e de expansão do *dataset*.

AGRADECIMENTOS

Os autores agradecem ao INSTITUTO FEDERAL DE SÃO PAULO e ao PROGRAMA DE APOIO À CIÊNCIA E TECNOLOGIA (PACTec) pelo apoio acadêmico e financeiro que contribuiu para o desenvolvimento deste trabalho.

REFERÊNCIAS

DALALAH, Doraid; DALALAH, Osama MA. The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. **The International Journal of Management Education**, v. 21, n. 2, p. 100822, 2023.

DOUGLAS, Michael R. Large language models. arXiv preprint arXiv:2307.05782, 2023.

KULKARNI, Neha; VAIDYA, Ravindra; BHATE, Manasi. A comparative study of word embedding techniques to extract features from text. **Turkish Journal of Computer and Mathematics Education**, v. 12, n. 12, p. 3550-3557, 2021.

LI, Ruiguang et al. A review of machine learning algorithms for text classification. **Cyber Security**, v. 1506, p. 226-234, 2022.

ODRI, Guillaume-Anthony; YOON, Diane Ji Yun. Detecting generative artificial intelligence in scientific articles: evasion techniques and implications for scientific integrity. **Orthopaedics & Traumatology: Surgery & Research**, v. 109, n. 8, p. 103706, 2023.

TANG, Ruixiang; CHUANG, Yu-Neng; HU, Xia. The science of detecting LLM-generated text. **Communications of the ACM**, v. 67, n. 4, p. 50-59, 2024.