Project Title: Detection of AI-Generated Content Across Various Contexts Using Classification Algorithms

Brian Cong

Introduction

The impact of the release of advanced language models, such as OpenAI’s GPT-3.5 and 4.0, Google’s PaLM, or Meta’s LLaMa, has created a critical need for the ability to differentiate AI generated text from human writing. In order to better address this need, I intend to further expand and explore areas of research involving AI text discrimination. How accurately can we discriminate between AI and human writing, what features should play significant roles in doing so, and how generalized can we make the detection approach?

Literature Review

The most recent existing solutions include specialized detectors such as the RoBERTa detector, (<https://arxiv.org/abs/2306.07401>) as well as various other algorithms available on the open market, such as Copyleaks AI detector, or GPTZero. One notable paper (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10328544/#R8>) outlines the use of XGBoost in order to detect AI generated text with a higher degree of accuracy within specifically medical academic writing using a targeted model, an area that other models, such as RoBERTa, struggle with. While the work therein is compelling, they note that their work has not been tested on other categories of academic writing, or other academic journals.

Methodology

Initial testing would involve first replicating the results found in the paper by Desaire et al., before attempting to expand generation and datasets using titles and paragraphs from other journals and categories of research. I would also propose to expand their work by fine-tuning their model further, exploring what kinds of inputs might increase accuracy — including document-level details might be a viable route for improving accuracy; details such as authorship, journalistic writing styles, subject matter, or other details may be able to improve model performance.

There is a wide body of work available for free from both medical journals as well as other science journals, and GPT-3.5 is available to generate AI-written journal entries for free. It may be worth doing a comparison of different AI models, such as GPT-4 or PaLM text generation, but the possibility of doing so will depend largely upon time and memory constraints. For classification I intend to use XGBoost much like the paper by Desaire et al., though I may compare and contrast model accuracy against RoBERTa and approaches.

Proposed Timeline

By milestone 2 I would expect to have a working adaptation of the original paper’s XGBoost implementation as well as some experimental data involving other journal entries and styles. By the final milestone I would expect to have results detailing the various possible inputs that could increase algorithmic accuracy across various subjects and journal-specific writing styles, as well as possibly being able to compare and contrast various other model accuracies and contexts.

References:

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing academic science writing from humans or CHATGPT with over 99% accuracy using off-the-Shelf Machine Learning Tools. *Cell Reports Physical Science*, *4*(6), 101426. https://doi.org/10.1016/j.xcrp.2023.101426

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). Tweepfake: About detecting deepfake tweets. *PLOS ONE*, *16*(5). https://doi.org/10.1371/journal.pone.0251415

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by CHATGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00819-6

Jawahar, G., Abdul-Mageed, M., & Lakshmanan, V.S., L. (2020). Automatic detection of machine generated text: A critical survey. *Proceedings of the 28th International Conference on Computational Linguistics*. https://doi.org/10.18653/v1/2020.coling-main.208

Wang, Z., Cheng, J., Cui, C., & Yu, C. (2023, June 9). *Implementing bert and fine-tuned Roberta to detect AI generated news by chatgpt*. arXiv.org. <https://arxiv.org/abs/2306.07401>

R. Kamath, A. Ghoshal, S. Eswaran and P. Honnavalli, "An Enhanced Context-based Emotion Detection Model using RoBERTa," *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2022, pp. 1-6, doi: 10.1109/CONECCT55679.2022.9865796.