University of Victoria
Department of Electrical and Computer Engineering
ECE 470: Artificial Intelligence
Spring 2024

Project Proposal

Date: 23 February 2024

Section: A01
Professor: Dr. Kin Fun Li
Team: The Vancouver Canucks
Uche Nwogbo - V00978185
Ryland Nezil - V00157326
Alex Spacek - V00974571

Problem Statement and Motivation

The problem identified and solved in this project is the identification of texts or paragraphs written by artificial intelligence (AI) software. As advancements in natural language processing and AI technologies continue to evolve, differentiating between human-generated and AI-generated text has become more difficult. This issue holds significance in various fields, including academic research, interpersonal communication and online content, where the originality of the content is crucial. Detecting AI-generated text is vital for ensuring transparency, credibility, and ethical use of AI-generated text. AI techniques are necessary to solve this problem due to the complexity and sophistication of modern AI models, which often mimic human writing styles and patterns. Traditional methods may fall short in accurately identifying AI-generated text, which makes it important to leverage advanced AI techniques to enhance the accuracy and efficiency of detection mechanisms, contributing to the integrity of information and maintaining trust in written communication.

Related Work

The problem of detecting machine-generated text has attracted significant attention since the first Large Language Models saw public release. OpenAI created the “GPT-2 Output Detector” alongside the release of GPT-2. This tool is based on the RoBERTa Natural Language Processor and, like ChatGPT itself, is a deep neural network model and was trained using the entirety of Wikipedia[1]. It achieved a 95% detection rate of GPT-2 generated text. However, it is not recommended for use on text samples of fewer than 50 words.
An alternative to this adversarial neural network approach is to perform empirical statistical analysis of the text to identify stylometric features. These techniques have been deployed with the goal of not only identifying machine-generated text, but tracing it back to the specific language model that authored it[2].

Problem Formulation

The objective function involves designing an algorithm to detect whether specific texts or paragraphs are generated by artificial intelligence. The objective is to use a comprehensive data set to write a program that differentiates between human and AI-generated content. The search space in this scenario comprises the various features, patterns, and linguistic nuances present in both human and AI-generated texts. It involves exploring combinations of parameters, such as sentence structure, vocabulary usage, and syntactic patterns, to identify distinguishing factors between the two. The size of the search space is substantial, given the intricacies of language and the sophisticated nature of AI models. To successfully navigate through this search space a neural network is created using large datasets that classify these text characterizations in both human and AI.

Programming Language

The coding tasks of this project will be completed using the Python programming language. Python is widely used in academia and industry for data science and machine learning applications. Libraries such as NumPy[3], Pandas[4], and Matplotlib[5] will enable rapid development of data handling, pre-processing, and visualization capabilities. In addition, the TensorFlow[6] framework will be used for its high performance machine learning algorithms. TensorFlow provides a wide range of tools for training and evaluating artificial intelligence models.
All of these programming tools are open-source.

Evaluation Approach

The solution will be evaluated primarily via inspection of its confusion matrix, which is a concise format for representing the accuracy of a machine learning algorithm. Confusion matrices serve to track all possible types of classification outcomes, for instance:

- true positive: algorithm classified AI-generated text as AI-generated text;
- false positive: algorithm classified human-written text as AI-generated text;
- true negative: algorithm classified human-written text as human-written text; and
- false negative: algorithm classified Ai-generated text as human-written text.

Analysis of a confusion matrix allows not only success rate quantification, but also insight in regards to wherefrom problems may be arising. Moreover, these results are indicative of where tweaks and changes can be made to improve the algorithm's performance margins.

Such an approach ideally involves two independent datasets, one for training and one for evaluation. Thanks to the booming popularity of generative AI, several large datasets (>1M samples) are available for free on the internet. Due to their enormous size, the best way to ensure dataset independence is to split one dataset into two parts. For example, a 1.4M sample dataset could be split into a 1M sample training set and a 400k sample evaluation set. This method ensures that no duplicates are present in either dataset, and thus that these datasets are as independent as is reasonably possible.

In summary, performance evaluation will be conducted through an iterative sequence of training, testing, and improving, where each testing instance produces a confusion matrix that can be used to inspire the next round of improvements.

Milestones

The primary milestones along with their expected completion times for this project are as follows:

1. initialize group development environment; **March 4 2024**
2. import master dataset; **March 8 2024**
3. implement neural network; **March 11 2024**

4. formalize training and evaluation procedures; **March 18 2024**
5. begin "train, evaluate, improve" iterations; and **March 25 2024**
6. complete final report. **April 4 2024**

Our intention is to use Google Colab for project development, as it provides a simple tool for concurrent development among team members. Major development milestones will be recorded in a GitHub repository. It has been decided that a strict division of labor will prove ineffective for this project. Instead, team members are expected to consistently build on each other's work within the Colab notebook, leaving comments and making Git commits as necessary. The team will conduct regular meetings to ensure all members are up to date with the status of the project and current priorities.

Bibliography

[1] “roberta-base-openai-detector” *huggingface.co*, Jan. 18, 2024. https://huggingface.co/openai-community/roberta-base-openai-detector (accessed Feb. 24, 2024).

[2] T. Kumarage and H. Liu, *Neural Authorship Attribution: Stylometric Analysis on Large Language Models*. 2023.

[3] “NumPy,” numpy.org. https://numpy.org/about/

[4] “pandas - Python Data Analysis Library,” pandas.pydata.org. https://pandas.pydata.org/about/

[5] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[6] TensorFlow, “Why TensorFlow,” TensorFlow. https://www.tensorflow.org/about