Detecting Machine Generated Text in a Sea of Content

Advanced NLP
Project Outline

Goodfellas

Pradhuman Tiwari: 2020115016
Hitesh Goel: 2020115003
Manav Chaudhary: 2021121003
Team Number 13

Motivation

Large Language Models (LLMs), such as ChatGPT and GPT-4, have transformed the landscape of content generation, making it challenging to distinguish between human-written and machine-generated text. By showcasing remarkable fluency and coherence in their text generation, these models have made it increasingly difficult to discern whether a piece of content was crafted by a human or generated by a machine.

This development, while awe-inspiring in its technological achievement, carries with it a sobering reality—a potent threat to the veracity of information and the integrity of communication. The very ability of these models that allows them to mimic human language and thought processes is a double-edged sword, as it not only opens up exciting possibilities for automation and content creation but also provides a powerful tool for the dissemination of misinformation and disinformation.

Since humans perform only slightly better than chance when classifying machine-generated vs. human-written text, there is a need to develop automatic systems to identify machine-generated text with the goal of mitigating its potential misuse.

Problem Statement

Through this project, we aim to address the growing concern of misinformation and content manipulation in the textual world by developing automatic systems for text classification, i.e., systems that can effectively discern between humans and machines.

We will focus on three subtasks:

Subtask A: Binary Human-Written vs. Machine-Generated Text Classification

Given a full text, determine whether it is human-written or machine-generated. Notably, this task encompasses multilingual aspects, with a specific emphasis on the detection of machine-generated text across five distinct languages. This will help in flagging potential misleading or biased information, considering the ubiquity of large language models across linguistic boundaries.

Subtask B: Multi-Way Machine-Generated Text Classification

Given a full text, determine who generated it. It can be human-written or generated by a specific language model. This enables us to track the influence of specific LLMs on generation.

Subtask C: Human-Machine Mixed Text Detection

Given a mixed text, where the first part is human-written and the second part is machine-generated, determine the boundary, where the change occurs. Through this, we wish to address cases where human and machine input combine. This will allow us to detect and mitigate potential manipulation within a text.

While our project initially aims to tackle all three subtasks, **our primary focus will be on the first two**. The extent of our project's scope expansion will depend on the resources, time, and effort available for implementation.

Additionally, it's worth noting that the dataset required for the third task has not been made available yet. However, we are optimistic that the dataset for this task will be released by the end of this month.

Dataset

We will explore available datasets for training and evaluation, including:

- LLM-generated text.
- Human-written text from various domains.
- Mixed text datasets.

The data for the task will be an extension of the M4 dataset. Training data for subtasks A and B can be found here. The following is a sample for each subtask:

subtaskA_monolingual

{"text":"1. Before applying for driving ... driving school,r...yadh (better than others)or any other driving school,...","label":1,"model":"bloomz","source":"wikihow"}
{"text":" Most card collectors ...in a way that makes logical sense.\n\n","label":0,"model":"human","source":"wikihow"}

subtaskA_multiligual

{"text":"Wer von den Macuils ... des Aztekenstaates.","label":1,"model":"chatGPT","source":"german"}
{"text":"فوق الأكسيد أو الأكسيد الف ... الأكسجين الكيميائية.","label":0,"model":"human","source":"arabic"}

- Our current exploration suggests that the multilingual dataset has issues that need resolution.
- Not all source languages are present in the train and validation datasets, and some entries seem wrong.
- We have mailed the organizers about the same, awaiting a response.

subtaskB

{"text":"Source Code (C++\/CUDA) for reproducing the results\n\n","model":"human","source":"peerread","label":0}
{"text":"The paper \"Exploring the Application of Deep Learning ... neural network architectures.","model":"chatGPT","source":"peerread","label":1}

The dataset for subtask C has not been released yet. If the stars align with our ambition, we shall overcome borders.

Literature Review

1. DetectGPT: The paper proposes a method that does not require training a separate classifier, collecting a dataset of real or generated passages, or explicitly watermarking generated text. Instead, it uses log probabilities computed by the model of interest and random perturbations of the passage from another generic pre-trained language model (e.g., T5). While rather inspiring, we believe the approach might not be viable for our tasks due to its reliance on log probabilities computed by the Language Model (LM). Nevertheless, we intend to explore it further as a potential source of inspiration within our study.
2. Machine-Generated Text: Claiming to be “the most complete review of machine-generated text detection methods to date”, this survey paper provides an overview of the methods used for machine-generated text detection, including feature-based approaches, neural language model (NLM)-based approaches, domain-specific research, human reviewers' abilities in identifying machine-generated text, evaluation methodology trends, and a discussion on prompt injection as a technique for shaping NLG model responses to aid in detection.
3. MGTBench: While we intend to use traditional metrics for evaluating the performance of our classification models, we intend to look at this newly proposed benchmark, which delves into a complex task similar to our Subtask B, involving the identification of the specific language model responsible for generating a piece of text. Our focus will be on examining the "LM Detector" method and its susceptibility to deceptive text alterations.
4. Detecting AI Content: This paper tests the accuracy of five AI content tools, GPTZero, OpenAI Text Classifier, Writer.com’s AI Content Detector, Copyleaks AI Content Detector, and Giant Language model Test Room, to detect ChatGPT, YouChat, and Chatsonic responses.
5. The Science of Detecting LLM-Generated Texts: A survey that provides an overview of existing LLM-generated text detection techniques, including an approach that uses inference time watermarks to detect text. A more thorough reading is required.

Since literature review is a continuous endeavor, we will persist in our search for relevant papers in this field.

We have identified the following **baselines and evaluation metrics** for our experiments:

- Metrics:
    - Accuracy, Precision, Recall, F1-score
    - Classification accuracy per LLM for Subtask B
- Baselines:
    - Pretrained BERT, GPT2

Timeline and Implementation Plan

Task	Objectives	Deadline
Literature Review, Dataset Exploration	• Includes exploratory analysis on the provided dataset. • Chalk out the implementation details, while exploring novel approaches from current literature.	September 19, 2023
Baseline Experiments	• Try out baselines on the available datasets. • Set up an evaluation pipeline. • Start implementing identified novel approaches.	October 2, 2023
Execution of Novel Ideas & Final Delivery	• Compile all baselines, evaluation scores and novel architectures into a fine package. • Report findings and suggest future possibilities. • Will include sufficient analysis of results, a detailed report and a well-designed, well-illustrated presentation.	November 2, 2023