Team 7: AI-generated Content/Fake News Detector

Names: Aryan Rakshit, Dhiya Pereira, Jack White, Seung Hyeon (Leah) Lee, Raj Penmetcha, Seth DeWhitt

Problem Statement

AI is very powerful, but that comes as a double edged sword. With few official regulations on its use, social media users of all age groups regularly struggle to identify AI-generated content. Most pre-existing AI-detecting tools function as a “copy-paste” system, relying on users to go out of their way to tab out of their social media platform in order to verify if text/images are AI-generated. In contrast, our extension offers in-site identification by providing automatic identification in the context of a feed, timeline or scroll.

Objectives

- Develop an extension that can detect AI in social media content in a quick, hands-off manner while the user is still browsing.
- Analyze incoming content (both text and images) using a lightweight AI detection system that distinguishes between human-made and AI-generated content in real time during normal speed scrolling. Preliminary goal of 1s identification time for text posts and 3s for image and video posts.
- Detect if the user input is text or image then use appropriate function/algorithm to generate result.
- Display a clear and concise result to the user. Display low, medium, or high probability for each post.
- Optimize the detection pipeline/algorithm so that the extension runs quickly enough to keep up with user scrolling, as a sluggish/bloated system would defeat the entire purpose.
- Seamless user-experience from before download -> regular use. Our project is geared towards audiences of all ages, ranging from tech-savvy teenagers to the elderly.
- Accept user feedback from diverse age groups to inform our design choices.

Stakeholders

1. Users : everyone who wants to check if the content is AI-generated, but do not want to disrupt their flow (i.e. opening up a new tab), tech-savvy youth with middle-aged parents or elderly grandparents
2. Developers : Aryan Rakshit, Dhiya Pereira, Jack White, Seung Hyeon (Leah) Lee, Raj Penmetcha, Seth DeWhitt

3. Project manager : Zhou Xuan
4. Project owners : Aryan Rakshit, Dhiya Pereira, Jack White, Seung Hyeon (Leah) Lee, Raj Penmetcha, Seth DeWhitt

Deliverables

1. AI Detection System: We will have our AI-generated content detection models for both text and image analysis displayed for use.
    a. For text analysis, we likely use some transformer-based model (ex. Hugging Face) and PyTorch. Similar to tools like GPTzero, we would have some website-based detector for manual checks and extend it with a Chrome extension interface
    b. For image analysis, likely CNNs and vision transformers, possibly trained on AI-generated vs real images, or an existing open source model created to identify AI generated watermarks, possibly SynthID or Stable Diffusion. Also Pytorch and OpenCV for general frameworks
    c. For our backend API, we can use FastAPI to hold model inferences to the website/chrome extension
2. Project Website: We aim to produce a public website that includes instructions on how to use our AI detector, an option to insert images and text into our product, and eventually additional features and Chrome extension installation instructions.
    a. Frontend: React, Tailwind CSS
    b. Backend: Same FastAPI endpoints as Chrome Extension
    c. Data Privacy Policy:
        i. Users' social media feeds must not be stored or used against them
        ii. Chrome Extension: Our end goal, once all initial plans for our model are completed, is to create a fully functional Chrome extension that automatically scans content for AI-generated text and images.
    d. JavaScript, Chrome Extensions API and HTML/CSS for UI overlays and warnings for potential AI or misinformation content.