DataHack 2023 Challenge

The Challenge

Your DataHack 2023 team is part of a data science group at a private equity firm. Leadership is evaluating an investment opportunity in a major e-commerce platform, but they are concerned about the growing issue of fake reviews, especially those generated by advanced language models. To address this concern, your team has been given access to a curated dataset of genuine Amazon reviews, fraudulent Amazon reviews, and text-davinci-003 generated reviews. Your firm's executives want to understand the following questions:

1. **How effectively can fake reviews be detected?**
2. **What role does user metadata and textual content play in the detection process?**
3. **What is the potential of LLMs in creating more sophisticated spam reviews in the future?**

Your task is to answer these questions and justify your answers using data. This challenge can be approached from multiple angles, such as:

- Utilizing natural language processing (NLP) to differentiate genuine and fake reviews
- Employing machine learning algorithms to analyze the importance of user metadata in detection
- Investigating the evolution and future implications of LLM-generated spam reviews
- Developing interactive dashboards or visualizations to communicate your findings

We highly encourage creativity! Feel free to tackle the challenge in any way, as long as it addresses the key questions.

Competition Deliverables & Evaluation

By the end of the competition, your team is expected to deliver **(1)** a concise 4.5-minute (or less) PowerPoint presentation about your idea and key findings, and **(2)** any code that you write/use. Judges will evaluate your work via 3 categories:

1. **Creativity**: How novel are your ideas? Is your submission exciting?

2. **Technical Mastery**: Is your proposal technically correct? Did you use any interesting technologies?
3. **Quality**: How effective is your presentation? How polished is your code? How well can you sell your idea?

Submissions should be sent to mlds.utexas@gmail.com with the subject line MLDS_datahack

Dataset Description

For your analysis, you are given a CSV: https://drive.google.com/file/d/1S8teNBz0I1eH6g3k4CIGi0Z06k0mxoMy/view?usp=sharing

This file contains the Amazon review data, which for each review includes the reviewer name, the review text, and the review time, along with other data. In addition to the original dataset, we have also generated additional fake samples using GPT-3 (specifically text-davinci-003), and added these to the data. This will allow you to analyze the best ways to detect fake reviews generated using the newest machine learning methods. We hope that these will provide an especially interesting challenge.

See the below table for dataset details:

Feature	Description
reviewerID	A unique identifier for the reviewer. This is a string that represents a specific person who wrote the review.
asin	A unique identifier for the product being reviewed.
reviewerName	The display name of the reviewer. This is a string containing the name that the reviewer chose to display publicly on their Amazon account.
helpful	An array of two integers representing the helpfulness of the review. The first integer is the number of helpful votes the review received, and the second integer is the total number of votes (helpful and unhelpful) the review received.
reviewText	The main text of the review. This is a string containing the full text that the reviewer wrote to describe their experience with the product.

overall	The overall rating of the product given by the reviewer. This is a float value ranging from 1.0 to 5.0, with 1.0 being the lowest rating and 5.0 being the highest rating.
summary	A brief summary of the review. This is a string containing a short title or description that summarizes the reviewer's opinion about the product.
unixReviewTime	The time when the review was written, represented as a Unix timestamp (the number of seconds since January 1, 1970). This is an integer value.
reviewTime	The time when the review was written, represented as a string in the format "MM DD, YYYY" (e.g., "01 30, 2014"). This is a more human-readable representation of the review time compared to the Unix timestamp.
llm	Whether this sample was generated by GPT-3 (as opposed to part of the original dataset).

Feel free to reach out to a mentor if anything is unclear—don't waste valuable time if you don't know something.

Online Resources Policy

Online resources are allowed, but you must cite your sources in the competition submission form (pasting the link is fine). Not doing so is plagiarism and will result in disqualification. Judges are familiar with what has/hasn't been posted online already, and they will be checking submissions to ensure compliance.