Project Proposal Title: **Social Media Toxicity Shield**

Author: Manasvi Goyal, manasvigoyal@g.harvard.edu

Data: This project will use Jigsaw Multilingual Toxic Comment Classification dataset, available at https://www.kaggle.com/datasets/julian3833/jigsaw-multilingual-toxic-comment-classification. This dataset contains a large collection of comments annotated for varying levels of toxicity.

Background and Motivation: Toxic content on online platforms is a growing issue, contributing to mental health challenges like anxiety and depression. With harmful posts impacting well-being, proactive measures are crucial. This project seeks to protect users through real-time toxicity detection and immediate mental health support, promoting healthier online interactions.

Problem: Manual moderation of toxic content is slow and reactive, relying heavily on user reports and delayed reviews. This allows harmful posts to linger, worsening mental health and fueling negativity. The goal is to develop an automated system that detects and categorizes toxic posts in real time and provides immediate access to tailored mental health resources.

Scope and Methods: A Language Model (BERT, GPT) will be fine-tuned on the dataset for multi-label toxicity classification in real time. The model will be integrated into a browser extension that monitors social media feeds, dynamically flags toxic content with visual warnings, and triggers immediate mental health support alerts.

Concerns & Limitations:

1. Ethical: Analyzing social media posts raises privacy concerns. Processing will occur locally whenever possible, and any server-side work will use anonymized data.
2. Data Bias: The Jigsaw dataset may contain inherent biases. Bias mitigation strategies will be implemented to ensure fair and accurate toxicity detection.
3. Computational: Integrating transformer models into a browser extension may pose performance challenges, so lightweight frameworks and optimizations will be explored.