Initial Project Proposal:
Toxic Comment Detection in Amazon Product Reviews

1. **Introduction and Problem Statement**

The internet is saturated with user-generated content, and platforms such as Amazon host hundreds of millions of product reviews written by consumers worldwide. While most reviews offer genuine product feedback, a small portion may contain toxic or offensive language that can harm other users, mislead buyers, or violate community standards. Detecting such content manually is infeasible due to the massive scale of the platform.

This project aims to address the problem of identifying toxic language in Amazon product reviews. Unlike traditional sentiment analysis tasks, which focus on classifying reviews as positive or negative, this project will focus on detecting toxicity, which includes insults, threats, profanity, and other forms of offensive language. Since the Amazon Reviews 2023 dataset lacks explicit toxicity labels, the project will involve creating a weakly supervised labeling strategy and developing a machine learning model capable of detecting toxic content in textual reviews.

The ultimate goal of this project is to build a classifier that distinguishes between toxic and non-toxic Amazon reviews, leveraging recent advances in natural language processing and transfer learning.

2. **Data Source**

The primary dataset used in this project is the Amazon Reviews 2023 dataset, available through Hugging Face: https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023. This dataset includes over 570 million reviews from more than 50 million users across 33 product categories. For example, each review in the dataset includes:

```
{'rating': 5.0,
 'title': 'Such a lovely scent but not overpowering.',
 'text': "This spray is really nice. It smells really good, goes on
really fine, and does the trick. I will say it feels like you need a lot
of it though to get the texture I want. I have a lot of hair, medium
thickness. I am comparing to other brands with yucky chemicals so I'm
gonna stick with this. Try it!",
 'images': [],
 'asin': 'B00YQ6X8EO',
 'parent_asin': 'B00YQ6X8EO',
 'user_id': 'AGKHLEW2SOWHNMFQIJGBECAF7INQ',
 'timestamp': 1588687728923,
 'helpful_vote': 0,
 'verified_purchase': True}
```

The reviews are primarily in English and reflect a diverse array of writing styles and topics, making the dataset suitable for training a generalized toxic comment classifier. Since toxicity is not explicitly labeled, we will generate labels using heuristic-based keyword matching and zero-shot predictions using pre-trained transformer models.

3. **Methods, Techniques, and Technologies**

The following high-level methods and technologies will be used throughout the project:

- *Data Preprocessing:*

    - Extracting and cleaning review text (removing HTML artifacts, lowercasing, filtering non-English content)
    - Sampling a balanced subset of reviews for binary classification (toxic vs. non-toxic)
- *Weak Supervision and Labeling:*
    - Using keyword heuristics (e.g., presence of profanity, slurs, or insults) to label a subset of reviews as toxic
    - Employing zero-shot classification using a pre-trained transformer (e.g., RoBERTa or GPT-3 via API) to validate or refine toxic labels
- *Model Training:*
    - Fine-tuning a pre-trained transformer model (e.g., DistilBERT, RoBERTa, or BERT) on the weakly labeled dataset for binary classification
    - Using transfer learning to adapt a toxicity classifier to the Amazon review domain
- *Evaluation:*
    - Model evaluation using metrics such as precision, recall, accuracy, and F1-score
    - Conducting qualitative analysis on model predictions to assess false positives and false negatives
- *Technologies:*
    - Python, Hugging Face Transformers, PyTorch or TensorFlow, scikit-learn, pandas, and Jupyter Notebooks

4. Products to Be Delivered

The following deliverables will constitute the final project submission and form the basis for grading:

- *Labeled Dataset:* A curated and preprocessed subset of Amazon reviews with inferred toxicity labels (weak supervision)
- *Trained Classification Model:* A fine-tuned transformer model capable of identifying toxic reviews
- *Evaluation Report:* A summary of model performance metrics, including confusion matrix, precision, recall, and F1-score
- *Code Repository:* A well-documented GitHub repository with all scripts for data preprocessing, labeling, training, and evaluation
- *Final Presentation:* A short presentation or poster summarizing the project problem, methodology, results, and insights

This project will demonstrate the effectiveness of NLP methods in content moderation tasks and explore the application of weak supervision techniques on large, unlabeled datasets.