Hackathon Challenge: AI for Safer Online Spaces for Women

Overview

On the occasion of **International Women's Day**, this hackathon aims to create AI-driven solutions that foster **safer, more inclusive, and respectful online conversations**. Social media platforms and online forums often become breeding grounds for **misogyny, toxicity, and harmful content**, discouraging meaningful discussions and creating **hostile environments, especially for women's groups**.

Participants will **leverage AI and NLP** to detect, classify, and explain harmful discourse while ensuring **context-awareness**. The ultimate goal is to **analyze online discussions, identify toxic patterns, reconstruct meaningful conversations, and promote healthier digital interactions**. This hackathon is an opportunity to **drive real change** in online conversations by **redefining how digital discussions are moderated, reconstructed, and analyzed**.

Challenge Tracks & Objectives

1. Parent-Child Conversation Reconstruction

- **Goal:** Rebuild threaded discussions from fragmented online conversations.
- **Tasks:**
  - Identify **conversation flow** and detect missing or misleading context.
  - Develop models to **summarize key discussion points** while preserving intent.

2. Subreddit-Based Topic Classification

- **Goal:** Categorize discussions based on their topics within different subreddits.
- **Tasks:**
  - Train a classifier to **predict subreddit categories** based on text content.
  - Use **NLP techniques** to analyze trends in topic distribution.
  - Develop an **interactive visualization** of topic evolution.

3. Detecting Toxic or Harmful Comments

🔹 **Goal:** Identify and mitigate **toxic**, **hateful**, or **harassing** comments.
🔹 **Tasks:**

- Develop an NLP model to classify comments as **toxic, neutral, or non-toxic**.
- Suggest **context-aware alternative phrasings** to encourage positive discourse.

4. Context-Aware Misogyny Detection

🔹 **Goal:** Build a model that detects **misogyny** in online conversations while considering **context**.
🔹 **Tasks:**

- Train a model to classify misogynistic language while **differentiating between sarcasm, jokes, and harmful intent**.
- Highlight **problematic words or phrases** to explain model predictions.
- Propose **AI-driven moderation tools** to **flag, warn, or educate users**.

Dataset:

The provided dataset contains Reddit discussions with a parent-child conversational structure, annotated for misogyny, toxicity, and context. It includes metadata such as subreddit, author, timestamps, and classification labels.
Reference paper: An Expert Annotated Dataset for the Detection of Online Misogyny.

Evaluation Metrics

Criteria	Weight	Description
Model Accuracy	40%	Classification performance on test data.
Context Awareness	20%	How well the model understands conversation flow and nuances.
Explainability	20%	How the model highlights problematic content.
Innovation	10%	Use of novel techniques, feature engineering, or visualization.
Impact & Usability	10%	Practicality for real-world applications.

Deliverables

To ensure a structured and evaluable submission, participants must provide:

Code Repository

- ☐ A **GitHub repository** (or similar platform) containing:
  Full source code, scripts, and dependencies
  Clear instructions for running the model
  Documentation on dataset preprocessing and model training

Model & Evaluation Report

- ☐ A report or presentation covering:
  Model architecture and methodology
  Evaluation metrics (accuracy, explainability, etc.)
  Challenges faced and solutions implemented

User Interface (Optional but Encouraged)

- ☐ A working **prototype** where users can:
  Input online text (tweets, comments, etc.)
  View sentiment/misogyny/toxicity classification results
  See highlighted words that influenced the model's decision

Judging Criteria

✔ **Ethical Considerations:** Is the model fair and unbiased?
✔ **Performance & Generalization:** Can it work across different discussions?
✔ **Potential for Deployment:** Could this solution be integrated into social platforms?
✔ **Innovation:** Is there a unique approach to the problem?

Bonus: Implement **bias mitigation** techniques to reduce false positives/negatives.