AI Safety Chatbot: A Safer Way to Chat Online

Author: Yash Raj Varun
Date: September 2025

1. Introduction

In today's digital world, chatting online is fun, fast, and convenient. But it can also be unsafe. Online conversations can sometimes include abusive language, distressing messages, or content that isn't suitable for younger users. These risks are especially concerning for children and teenagers who are still learning how to navigate online interactions.

Our project, the **AI Safety Chatbot**, is designed to make online chats safer. It acts as an intelligent moderator that can detect harmful, inappropriate, or concerning messages in real-time and respond appropriately. The chatbot is deployed using Google Colab and Gradio, making it accessible and easy to share.

The core idea is simple: **help users communicate safely, while alerting them—or moderators—when there's potential danger.**

2. Project Overview

The AI Safety Chatbot works by combining **artificial intelligence, sentiment analysis, and user-friendly design**. Here's how it works:

Detection Modules

The chatbot has four main modules that evaluate messages:

1. **Abuse Detection** – Uses a pre-trained BERT model (unitary/toxic-bert) to detect toxic or abusive language. For example, if a user writes an offensive comment, the bot flags it immediately.
2. **Escalation Detection** – Monitors the overall mood of the conversation. If the conversation starts trending negative, it triggers a warning, helping moderators notice when things may escalate.
3. **Crisis Detection** – Scans messages for signs of self-harm or distress using keywords like "suicide" or "I want to hurt myself." This module helps identify urgent situations where human intervention may be needed.
4. **Age-Appropriate Filtering** – Blocks messages containing adult topics, such as drugs, sex, violence, or gambling, for users under 13. This ensures younger users aren't exposed to inappropriate content.

Chatbot Logic

All detection modules are combined in a single function that decides how to respond to user messages. The chatbot can output:

- ✓ Safe message processed
- 🚨 Abusive language detected
- ⚡ Escalation detected
- ⚠ Crisis detected! Escalating to human intervention
- 🔒 Content blocked for age-inappropriate messages

Each message is stored in a **persistent conversation file**, meaning the chat history is saved even if the notebook is reloaded.

User Interface

We built a **simple and interactive interface** using Gradio. Users can type messages, enter their age, and see responses in real-time. Previous messages are displayed, so the conversation feels continuous. The interface also generates a public URL, allowing anyone to interact with the chatbot from a browser.

3. Sample Interaction

Here's a glimpse of how the chatbot handles messages:

User Message	**Chatbot Response**
"I love coding!"	✓ Safe message processed.
"You are stupid!"	🚨 Abusive language detected.
"I feel worse than yesterday."	⚡ Escalation detected.
"I want to hurt myself."	⚠ Crisis detected! Escalating to human intervention.
"Tell me about drugs." (age 12)	🔒 Content blocked (age-inappropriate).

This demonstrates how the chatbot responds appropriately to various types of messages, balancing safety with usability.

4. Technical Highlights

- **Pre-trained AI Model:** Efficient and ready-to-use toxicity detection with BERT.
- **Real-Time Processing:** Messages are evaluated instantly.
- **Persistent Storage:** Chat history is saved for continuity.
- **Flexible Deployment:** Runs in Google Colab with public URL sharing.

These features make the AI Safety Chatbot an effective tool for safer online communication.

5. Impact and Future Work

The AI Safety Chatbot is more than a technical project—it's a step toward **responsible digital interactions**. It can be used in classrooms, social platforms, or private chat applications to protect users from harmful content.

Future improvements could include:

- Notifications for crisis alerts via email or messaging platforms.
- Multi-lingual support for global audiences.
- Integration with mobile and web apps for real-world deployment.
- More sophisticated sentiment analysis to detect subtle emotional cues.

6. Conclusion

In conclusion, the AI Safety Chatbot demonstrates how **artificial intelligence can make online communication safer, more responsible, and more human-friendly**. By combining abuse detection, crisis monitoring, sentiment tracking, and age-appropriate filtering, this project ensures that online conversations can be engaging without compromising user safety.

It's a simple idea with **big potential impact**, and it's designed to be shared, tested, and improved over time.

"AI shouldn't just make life easier—it should make it safer too."