COMMENT TOXICITY DETECTION SYSTEM

To determine the level and category of toxicity of comments in public forums .

Guided by
Er. Shyam Maheshwari

Submitted by
Akshat Bhatt (21C7005)
Ayan Nema(2117018)
Kamakshi Agrawal(2217093)
Siddhi Chouhan(2117077)

AGENDA

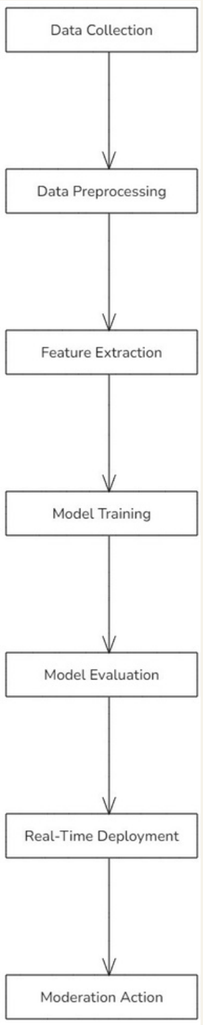
- Problem Statement
- Solution Overview
- Functionalities
- Methodology
- Project Overview
- Use Case Diagram
- Sequence Diagram
- Implementation And Testing
- Software Requirement
- Future Expansions

PROBLEM STATEMENT

The "Comment Toxicity" project addresses the rising issue of harmful online comments, such as hate speech and bullying, which impact users' mental health and create hostile digital environments. The goal is to develop an automated system using machine learning and natural language processing (NLP) to detect and moderate toxic content across platforms, improving user safety and experience.

SOLUTION OVERVIEW

The "Comment Toxicity" project proposes a machine learning-based solution to automate the detection and filtering of toxic comments. This solution leverages Natural Language Processing (NLP) and deep learning models to accurately classify comments as toxic or non-toxic. By utilizing a pre-trained language model, LSTM-based neural network, the project can analyze language patterns, tone, and context within comments, achieving a high level of precision in detecting toxicity.



Flow Chart Diagram

FUNCTIONALITIES

Automated Toxicity Detection

The core functionality is an automated comment classification system that identifies toxic content in real-time.

Multi-language Support

Adaptation to detect toxic comments across various languages, dialects, and slang to increase the model's applicability on a global scale.

Context Awareness

Use of NLP to understand context, reducing the likelihood of misclassifying ambiguous comments (e.g., sarcasm or idiomatic expressions).

Reporting and Analytics

Track flagged content statistics, monitor false positive/negative rates, and generate reports to help moderators optimize platform guidelines.

Methodology

1. **Data Collection and Preprocessing:**

   The project involves gathering a dataset of online comments labeled fortoxicity levels,typically from open-source datasets like the Kaggle Comment Classification dataset.

2. **Feature Extraction and Text Representation**

   To represent the text data numerically for model input,techniques like Tokenization and vector embedding were used.

3. **Model Selection and Training**

   Several machine learning models were tested to determine the most effective approach fortoxicity detection. Classical models, such as Support Vector Machines (SVM) and Naive Bayes, were evaluated alongside deep learning models like LSTM (Long Short-Term Memory networks) and transformer model.

4. **Model Serialization**

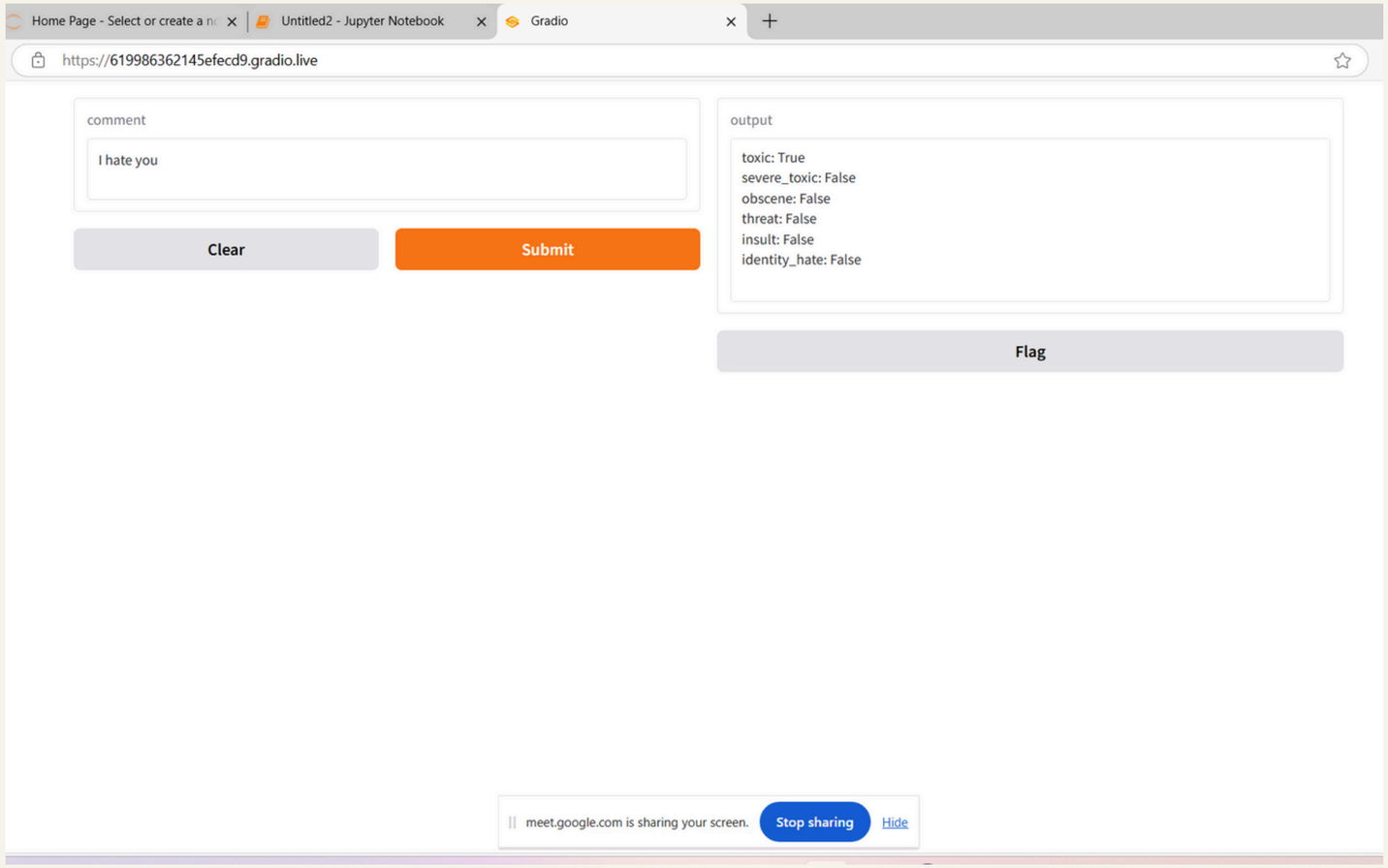
   Serialize the trained model and save it in an h5 format for ease of deployment and storage.

5. **Model Integration to UI**

   Integrate the serialized model into a Gradio app to create an interactive user interface (UI) that allows real-time testing of toxicity detection. This UI provides an accessible way for users to input comments and receive toxicity analysis results.

Project Overview

A user-friendly, dynamic portal is offered by this comment toxicity detection system to determine whether a comment is toxic, severe_toxic, obscene, insult, threat, or identity_hate.



comment

गृह मंत्री अमित शाह को बोन कैंसर से जूझ रहे हैं

Clear | Submit

output

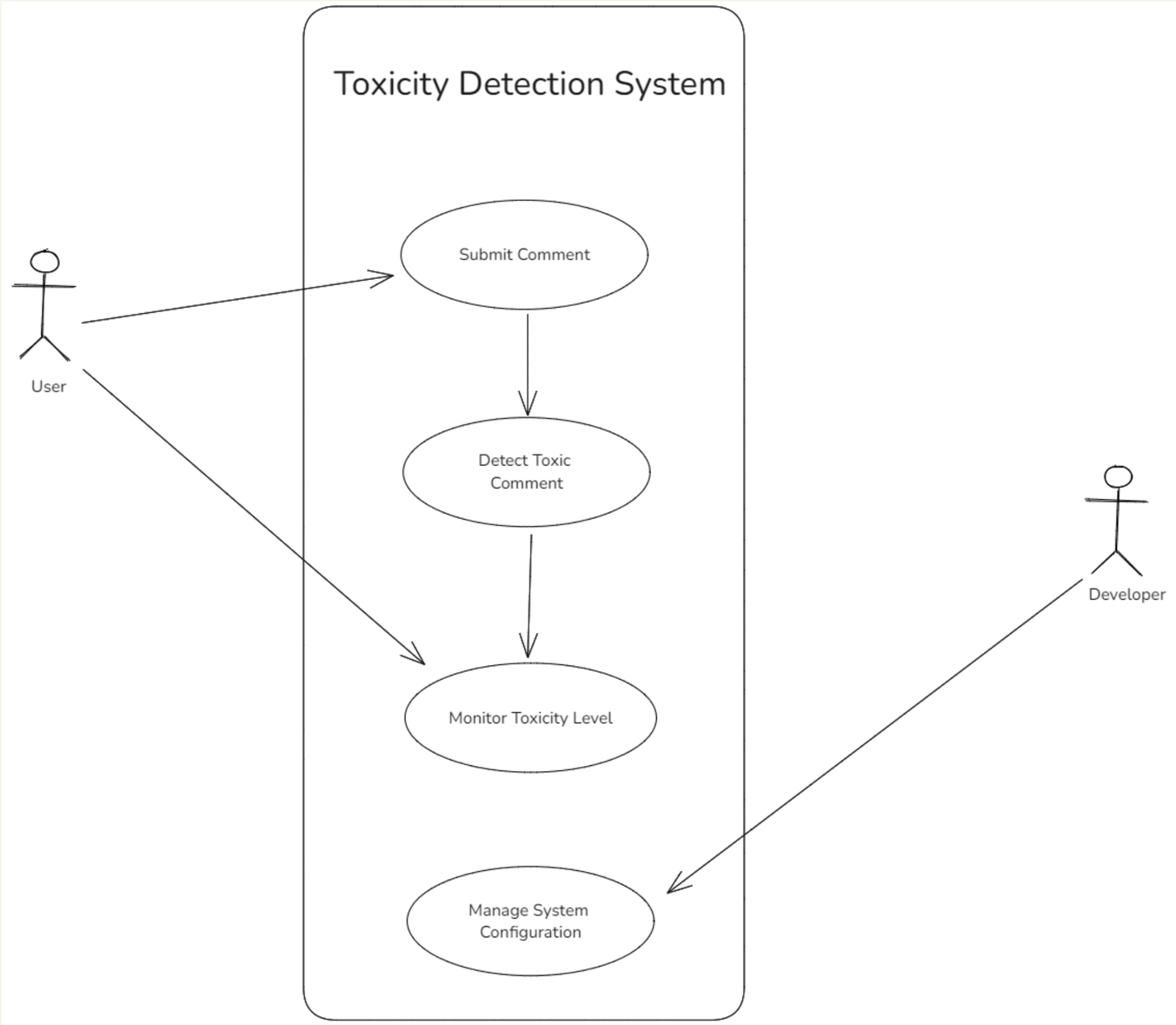
toxic: False
severe_toxic: False
obscene: False
threat: True
insult: True
identity_hate: True

Flag

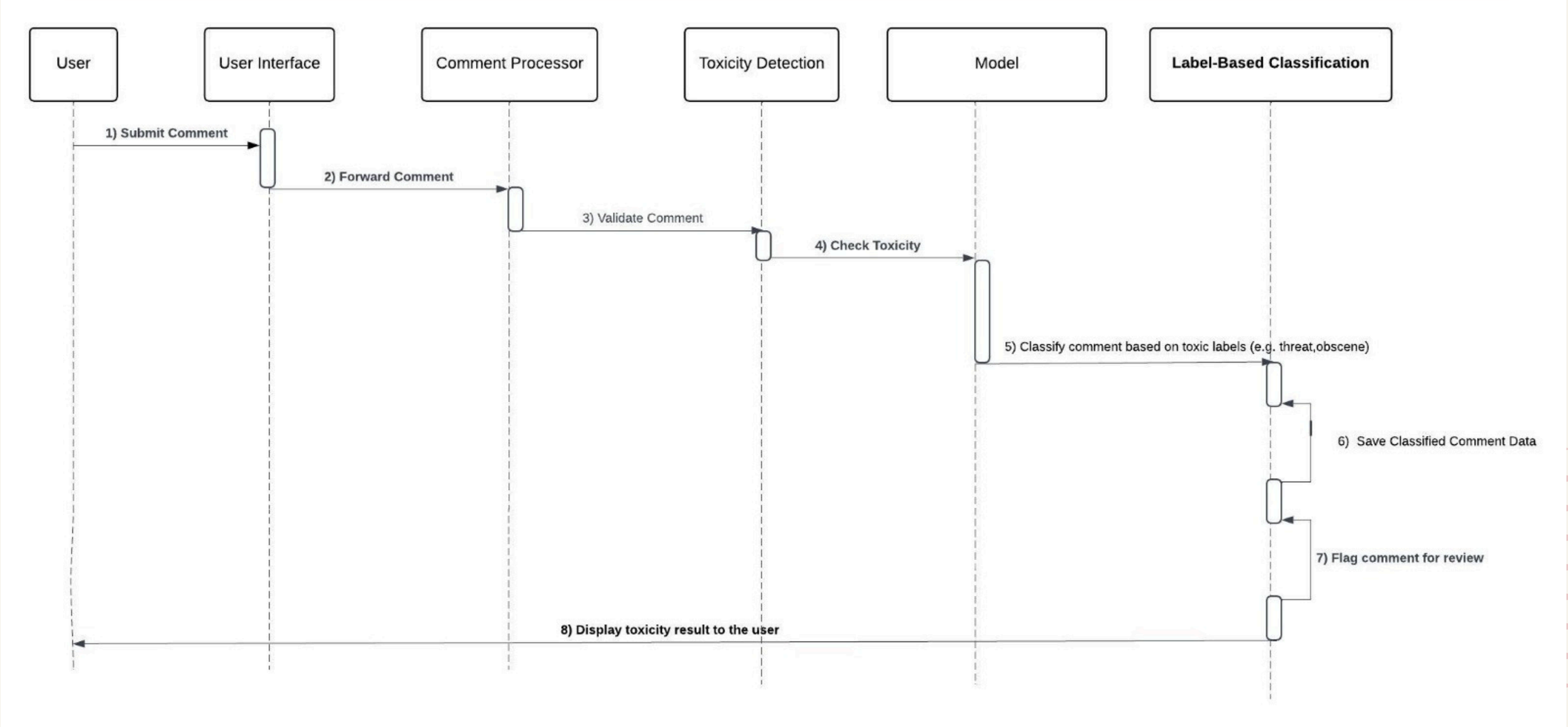
Use via API · Built with Gradio

Out[51]:

Use Case Diagram

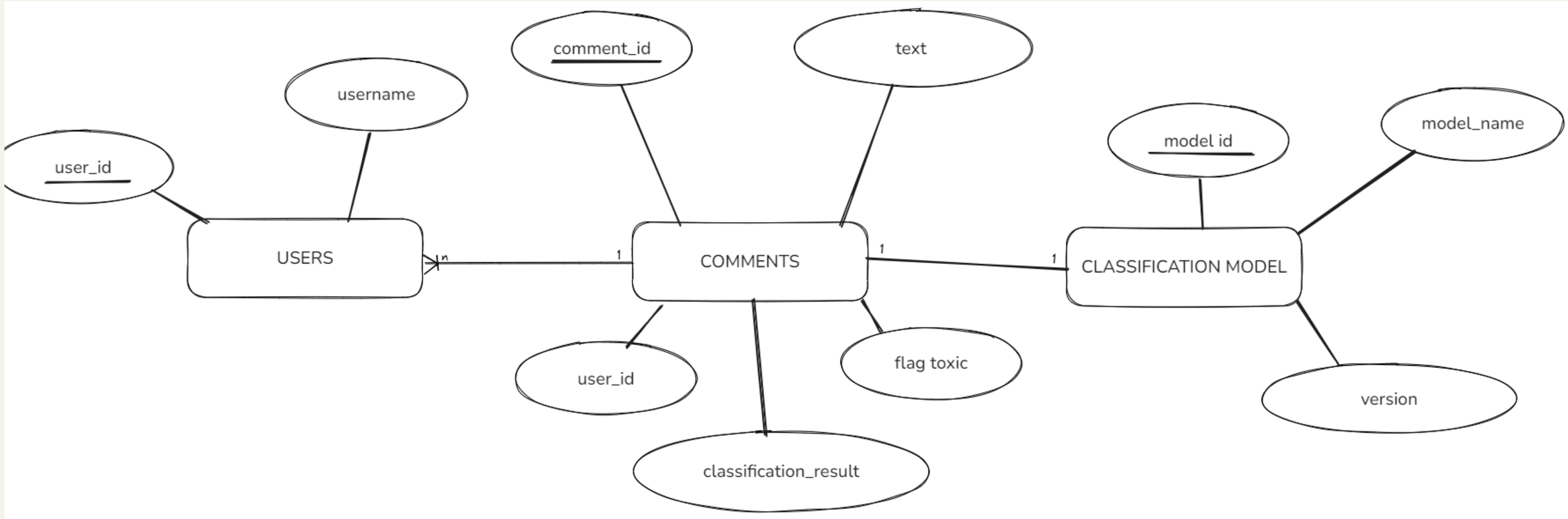


Sequence Diagram



Implementation And Testing

ER Diagram



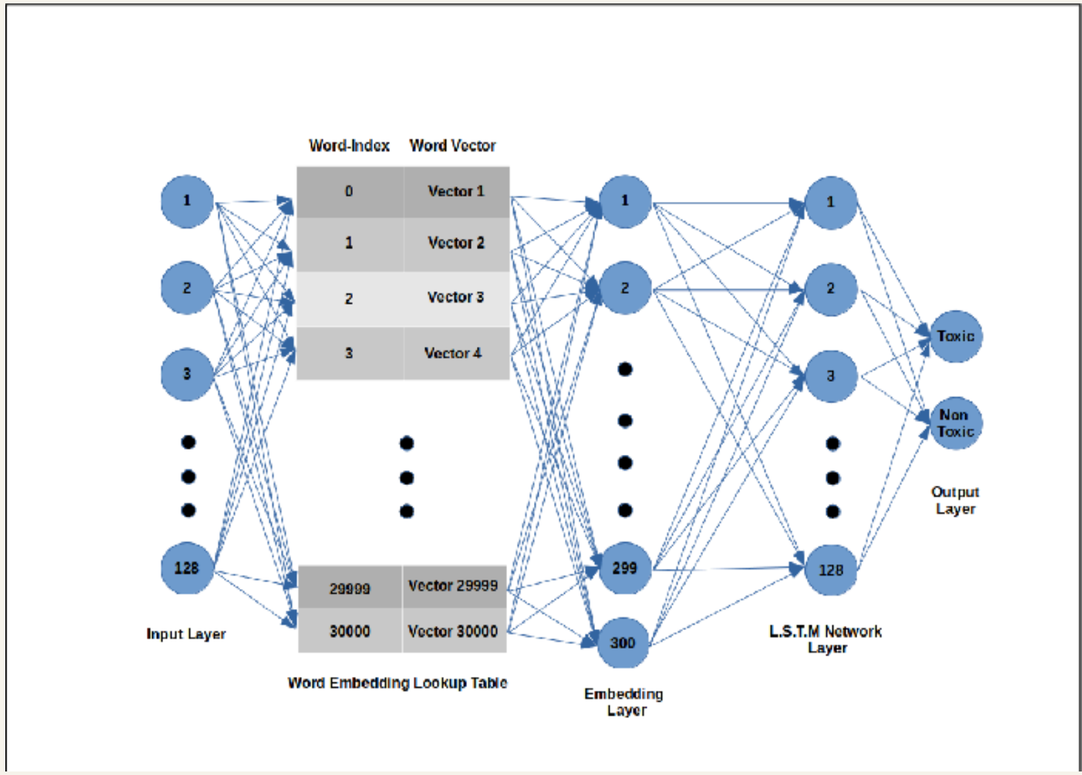
1. **Input Collection:**
The project involves gathering a dataset of online comments labeled for toxicity levels, typically from open-source datasets like the Kaggle Comment Classification dataset.

2. **Label Assignment**
Assign labels to these inputs that are multi-output and multi-binary (e.g., moderately toxic, severely toxic, threat, obscene etc).

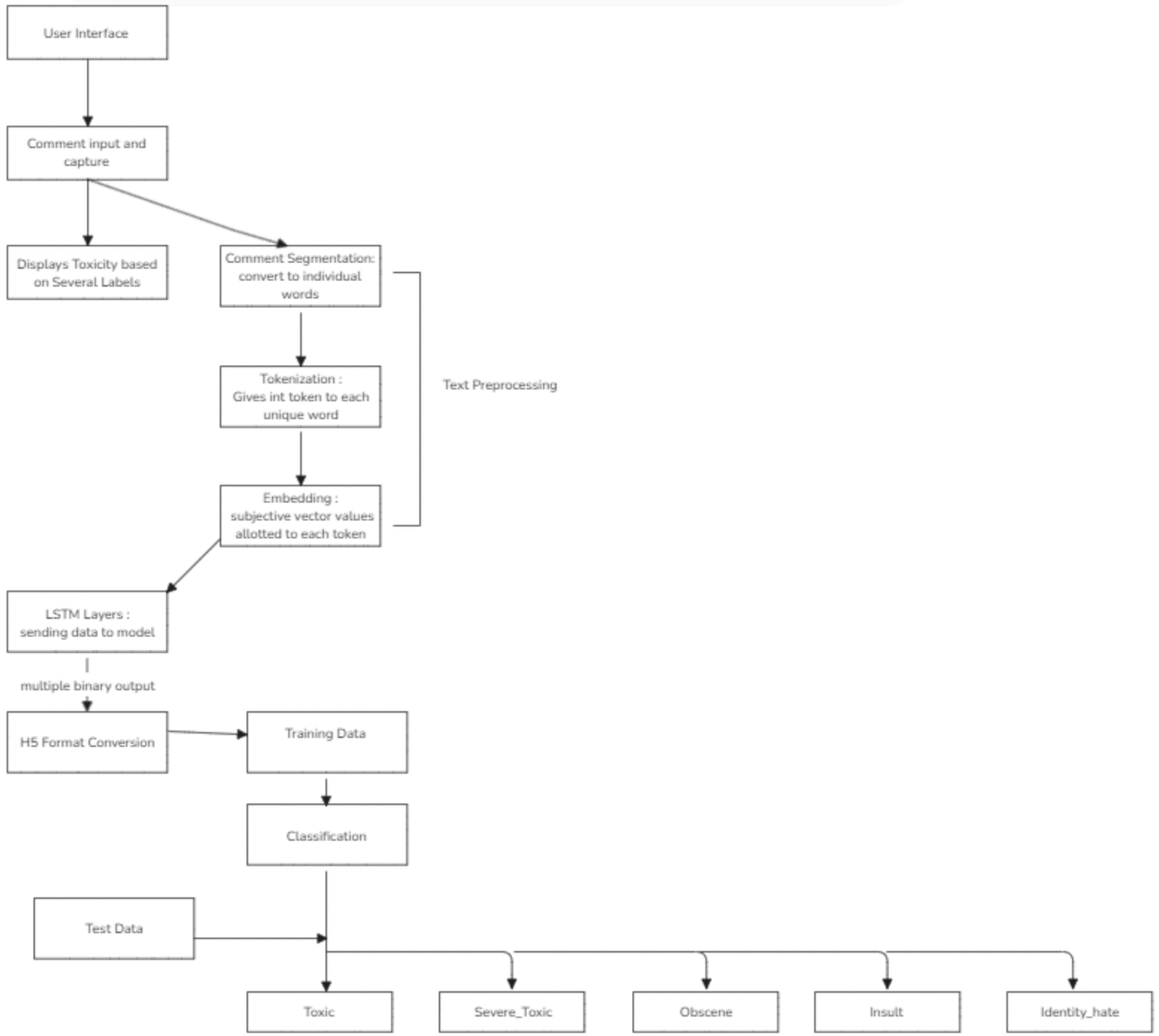
3. **Tokenization and Embedding**
Implement labels using tokenization, where each word is assigned an integer. LSTM layers are applied to process these sequences effectively.

4. **Model Serialization**
Serialize the model and save it in h5 format for easy storage and retrieval.

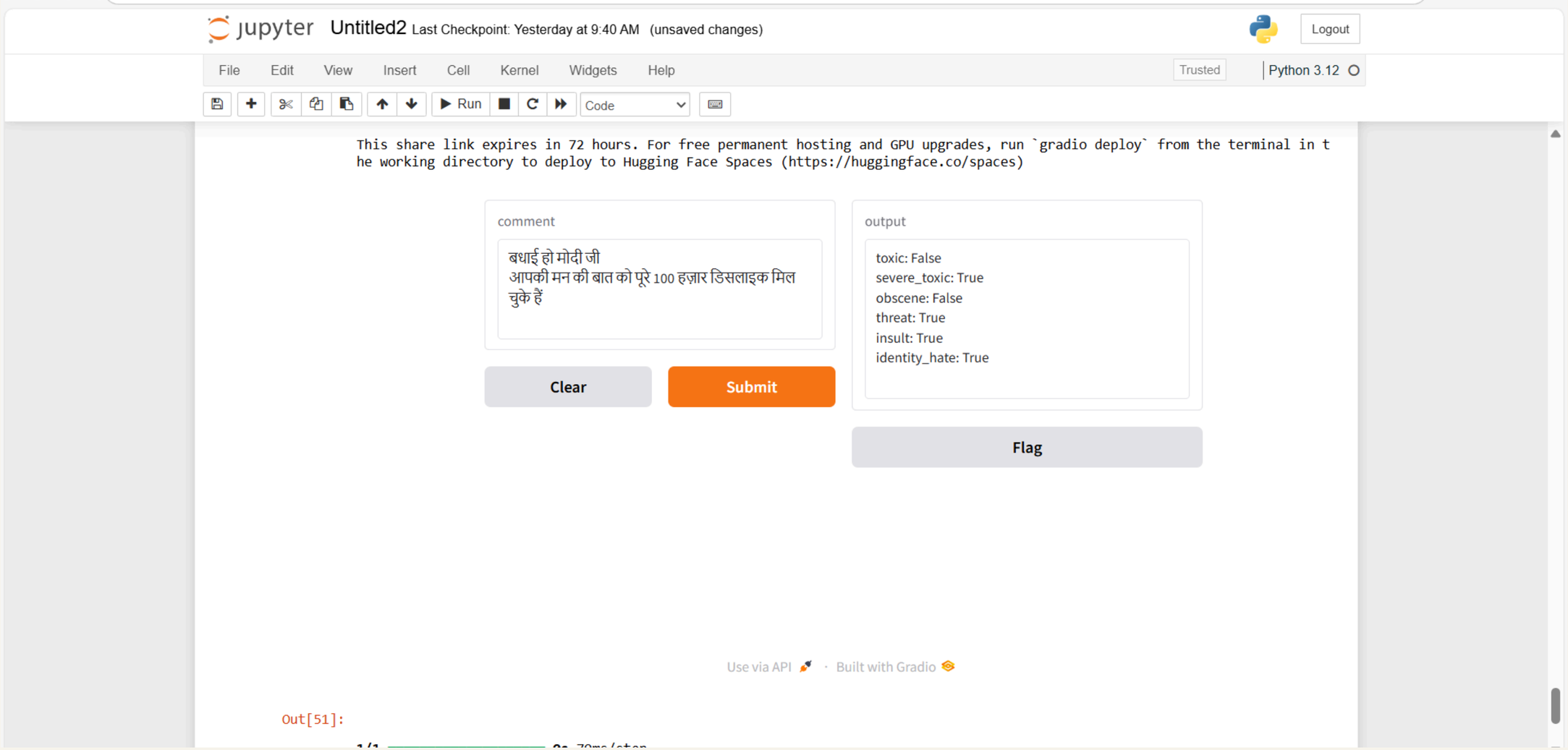
5. **Integration with Gradio App**
Integrate the h5 format model into a Gradio app to create a real-time testing UI for the application.



Architecture Diagram



Testing


jupyter Untitled2 Last Checkpoint: Yesterday at 9:40 AM (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3.12
Run
Code
This share link expires in 72 hours. For free permanent hosting and GPU upgrades, run `gradio deploy` from the terminal in the working directory to deploy to Hugging Face Spaces (https://huggingface.co/spaces)
comment
बधाई हो मोदी जी
आपकी मन की बात को पूरे 100 हज़ार डिसलाइक मिल चुके हैं
output
toxic: False
severe_toxic: True
obscene: False
threat: True
insult: True
identity_hate: True
Clear
Submit
Flag
Use via API · Built with Gradio
Out[51]:

comment

बधाई हो मोदी जी
आपकी मन की बात को पूरे 100 हज़ार डिसलाइक मिल चुके हैं

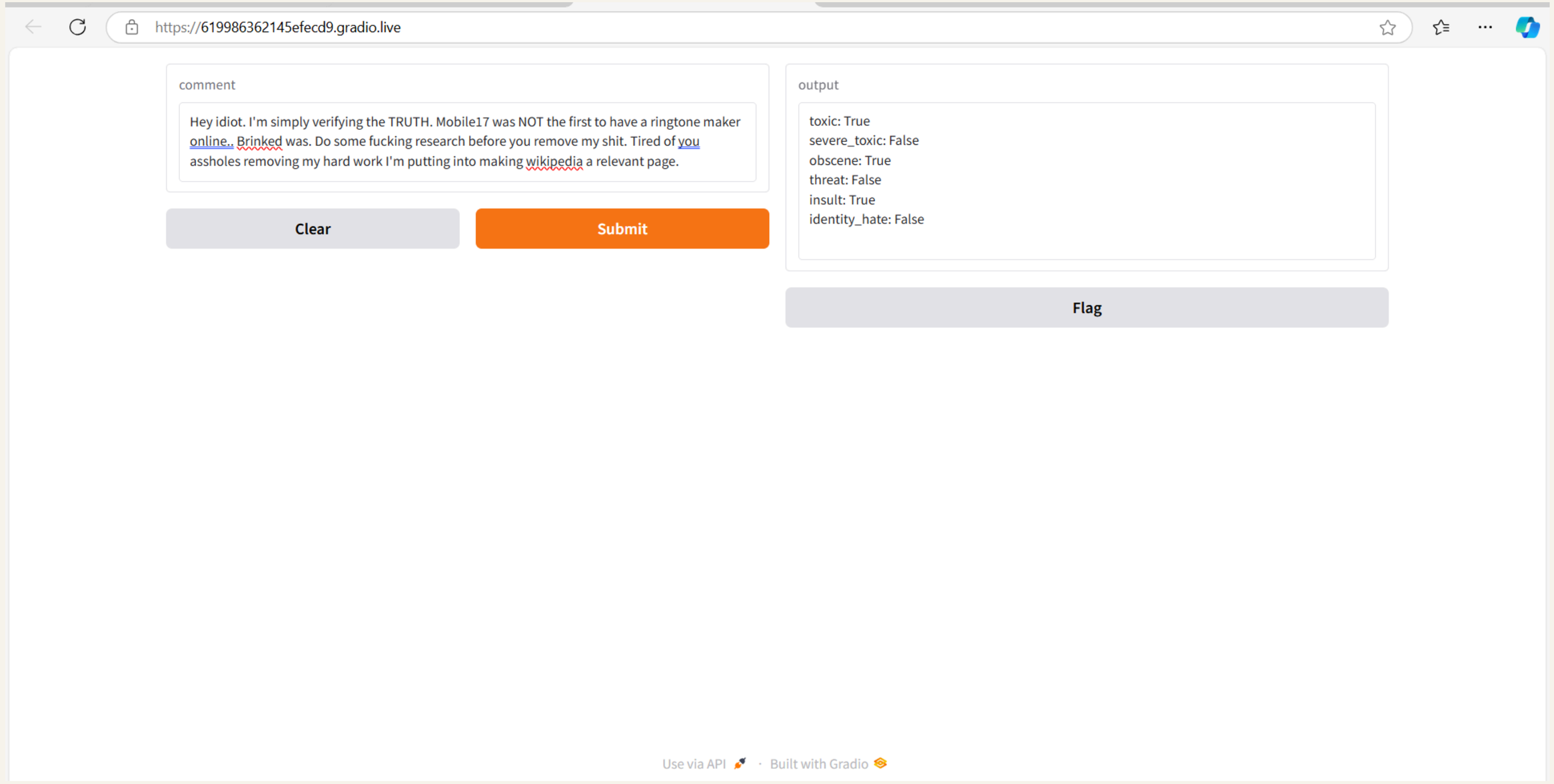
Clear

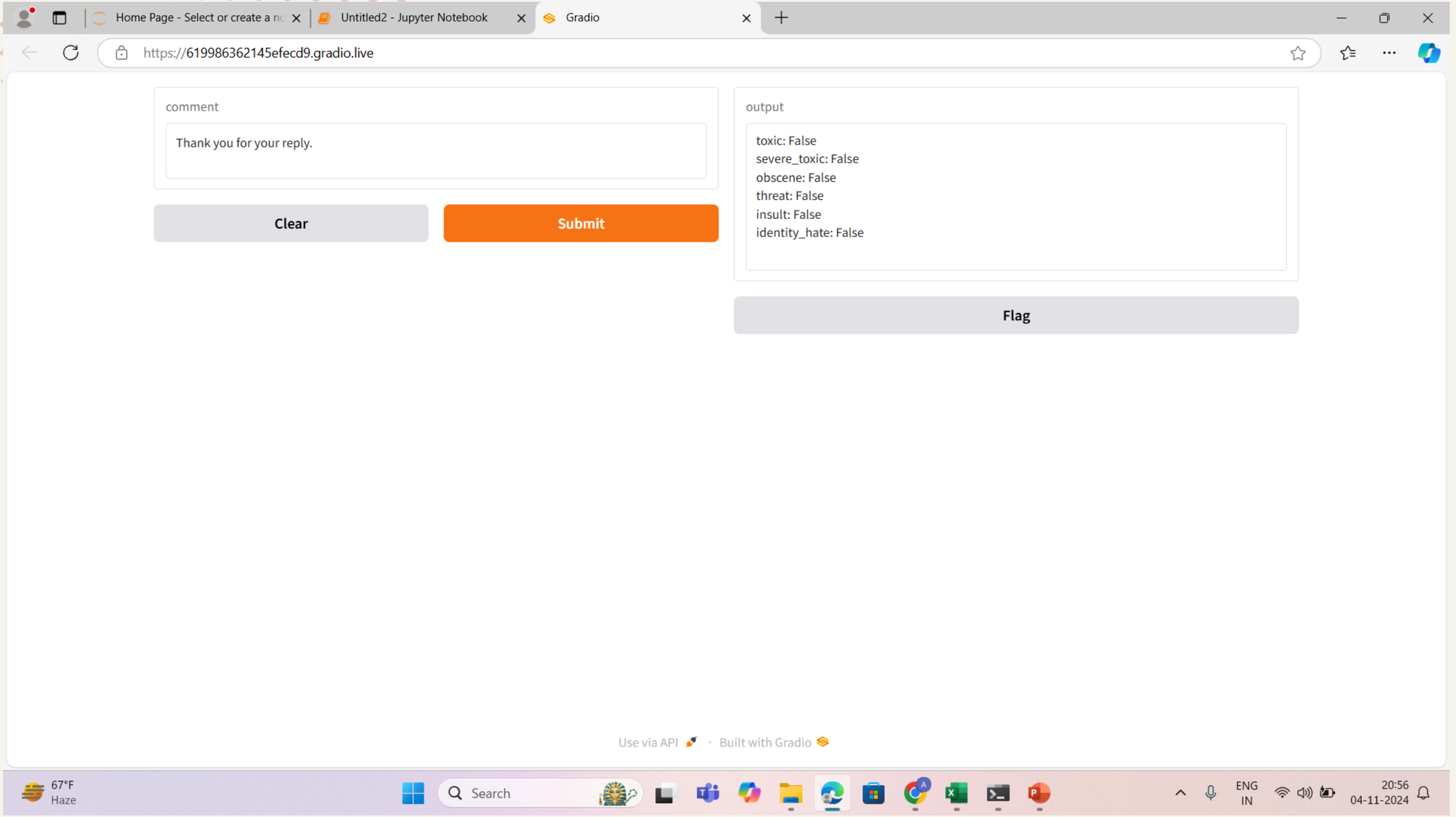
Submit

output

toxic: False
severe_toxic: True
obscene: False
threat: True
insult: True
identity_hate: True

Flag


https://619986362145efecd9.gradio.live
comment
Hey idiot. I'm simply verifying the TRUTH. Mobile17 was NOT the first to have a ringtone maker online.. Brinked was. Do some fucking research before you remove my shit. Tired of you assholes removing my hard work I'm putting into making wikipedia a relevant page.
Clear
Submit
output
toxic: True
severe_toxic: False
obscene: True
threat: False
insult: True
identity_hate: False
Flag
Use via API
Built with Gradio


Home Page - Select or create a n
Untitled2 - Jupyter Notebook
Gradio
https://619986362145efecd9.gradio.live
comment
Thank you for your reply.
Clear
Submit
output
toxic: False
severe_toxic: False
obscene: False
threat: False
insult: False
identity_hate: False
Flag
Use via API
Built with Gradio
67°F
Haze
Search
ENG
IN
20:56
04-11-2024

comment

You truly are the worst admin

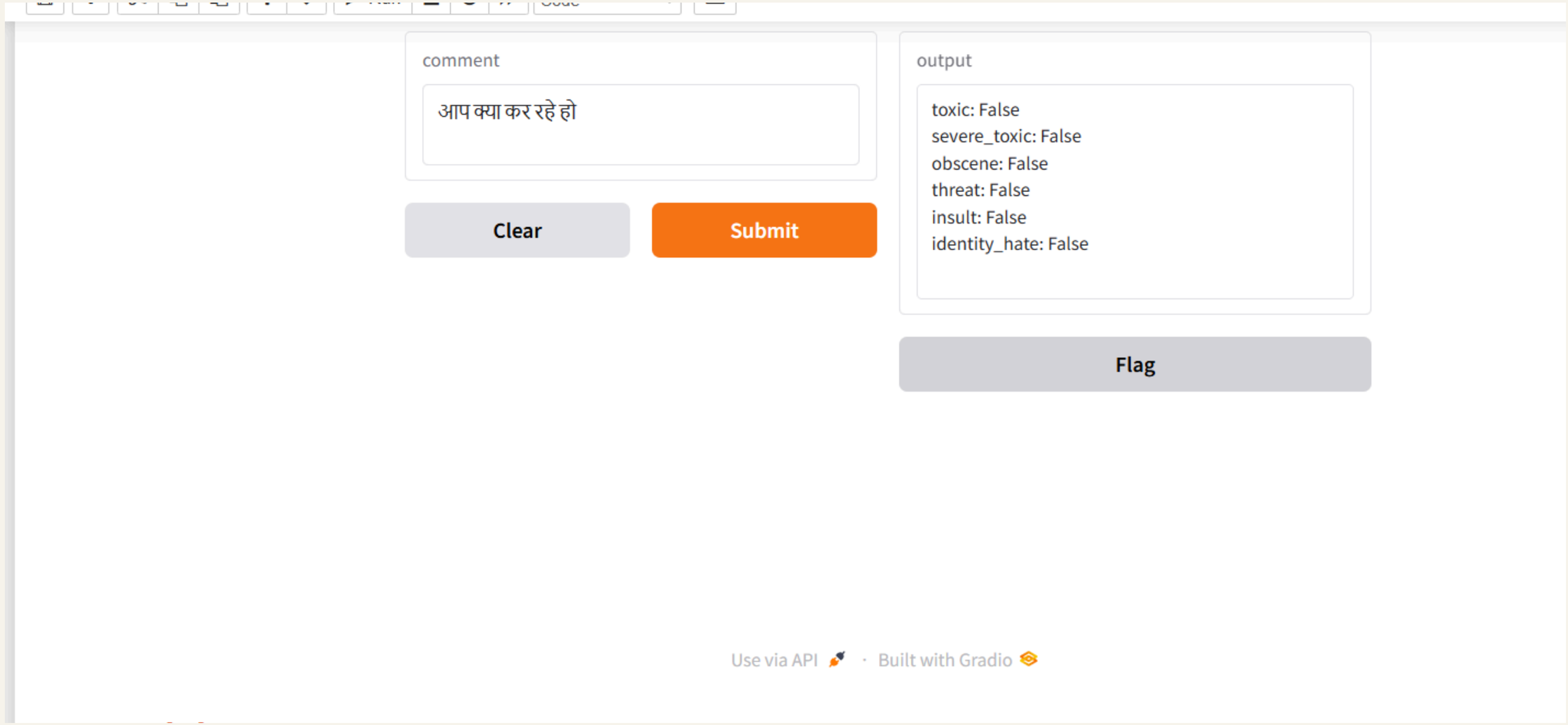
Clear

Submit

output

toxic: True
severe_toxic: False
obscene: False
threat: False
insult: False
identity_hate: False

Flag


comment
आप क्या कर रहे हो
Clear
Submit
output
toxic: False
severe_toxic: False
obscene: False
threat: False
insult: False
identity_hate: False
Flag
Use via API
· Built with Gradio

Software Requirements

- **Python: A versatile language with extensive libraries (e.g., TensorFlow, Keras) idea for NLP and deep learning, accelerating development.**
- **TensorFlow & Keras: Machine learning frameworks for efficiently building, training, and optimizing deep learning models, especially for text data in toxicity detection.**
- **Jupyter Notebook: An interactive environment for code, visualization, and documentation, ideal for iterative development in NLP projects.**
- **Pandas & NumPy: Libraries for efficient data manipulation and numerical computing, essential for preprocessing and dataset management.**
- **Gradio: A Python library for building quick web interfaces, enabling users to input comments and receive real-time toxicity feedback.**

Future Expansions

- **Improved Model Performance: Use advanced models like GPT or T5 and combine different approaches to improve understanding of complex toxic content and reduce errors.**
- **Larger and Fairer Datasets: Collect more diverse data and apply debiasing techniques to make the model more fair and adaptable to different cultures and languages.**
- **Sarcasm and Context Detection: Add sentiment analysis or context-aware layers to help the model detect sarcasm or context-specific toxicity. Using data like images or emojis can boost accuracy.**
- **Real-Time API and Feedback: Build a real-time API for wider use (e.g., social media), and use user feedback to improve model accuracy over time.**
- **Explainability: Include tools like LIME or SHAP to explain why a comment is flagged as toxic, improving transparency and trust.**

Thank You