CYBERBULLYING DETECTION

Navigating the Digital World Responsibly

PROJECT OBJECTIVES

- Problem Statement
- Existing System
- Proposed System
- End Users
- What Makes Our Project Unique
- Why BiLSTM?
- Modelling/Block Diagram/Flow of Project
- Screenshots
- Result/outcomes
- Conclusion
- Future Perspective

PROBLEM STATEMENT

In today’s world, most people use social media apps like Instagram, WhatsApp, and Telegram to connect with friends, share content, and communicate. But along with the good side, there's also a dark side – many users face cyberbullying in the form of abusive messages, toxic comments, or online threats. This often goes unnoticed and can seriously affect a person's mental health. There’s a need for an intelligent system that can automatically detect and flag such content to make online platforms safer.

Through this project, the aim is to develop a model that can detect cyberbullying in real-time and integrate it into social media-like platforms to prevent harmful communication.

EXISTING SYSTEM

Current cyberbullying detection methods utilize keyword filtering, Support Vector Machines (SVMs), and Recurrent Neural Networks (RNNs) to identify toxic content. While deep learning approaches such as BERT and GPT-based models have improved accuracy, they are computationally expensive and often lack scalability for real-time deployment. Moreover, existing solutions fail to capture nuanced patterns of cyberbullying, especially in multi-label toxic comment classification

PROPOSED SYSTEM

The proposed system leverages deep learning-based Natural Language Processing (NLP) techniques, specifically Bidirectional LSTM (BiLSTM) networks, to classify and detect harmful communication (e.g., toxic, severely toxic, obscene, threat, insult, and identity hate) in real-time. The system:

1. Trains on the Jigsaw Toxic Comment Classification dataset to learn complex linguistic patterns associated with cyberbullying.
2. Performs multi-label classification to categorize different types of harmful communication with high accuracy.
3. Utilizes a scalable and efficient deep learning model optimized for real-time processing.
4. Deploys via Gradio for an intuitive web-based interface, allowing instant detection and feedback.

This system offers a practical, scalable, and real-time approach to combating cyberbullying, enhancing online user safety, and reducing psychological distress caused by toxic interactions.

END USER

- Social media users exposed to online bullying
- Developers building safer chat or comment platforms
- Schools and colleges monitoring student interactions
- Parents ensuring online safety for their children

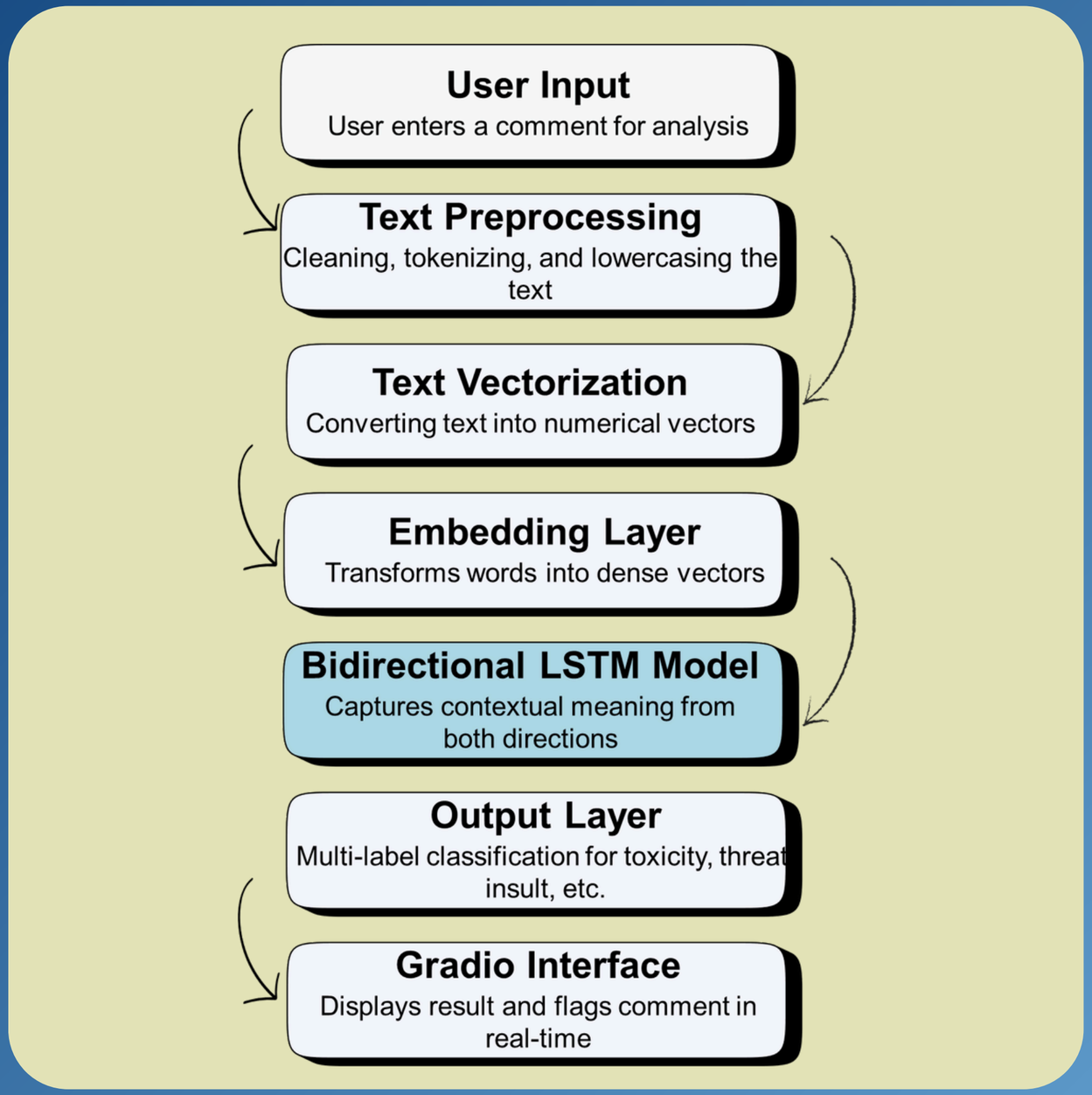
WHAT MAKES OUR PROJECT UNIQUE

- Uses a Bidirectional LSTM model for deep understanding of text context.
- Performs multi-label classification on toxic comment types like threat, insult, hate, etc.
- Real-time detection of cyberbullying via a web interface using Gradio.
- Trained on the large Jigsaw Toxic Comment dataset for better accuracy.
- Integrated into an Instagram-like UI for real-world simulation and demo.

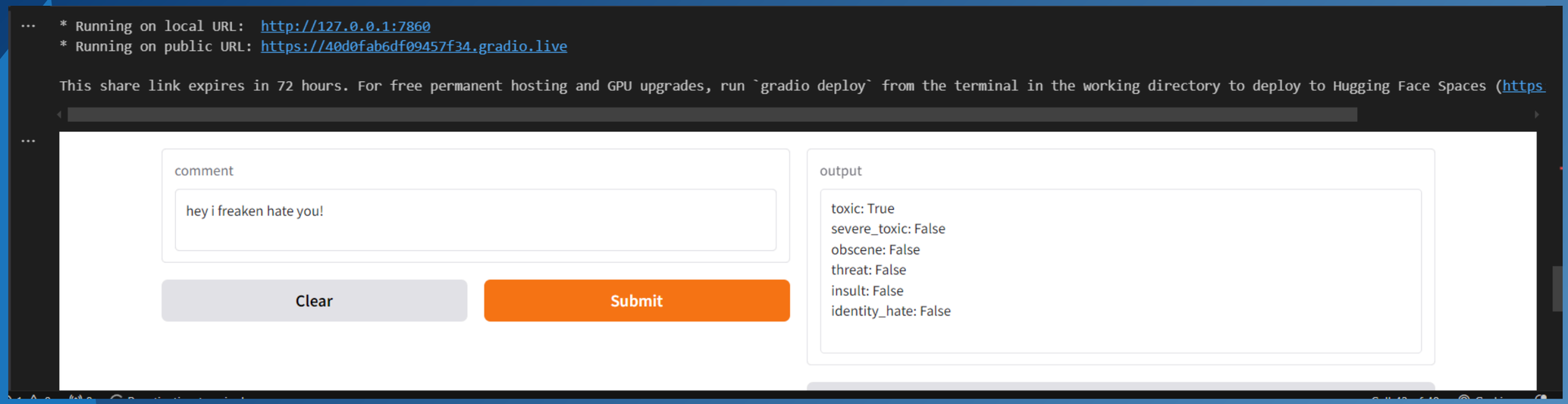
WHY BILSTM?
THE POWER OF UNDERSTANDING CONTEXT

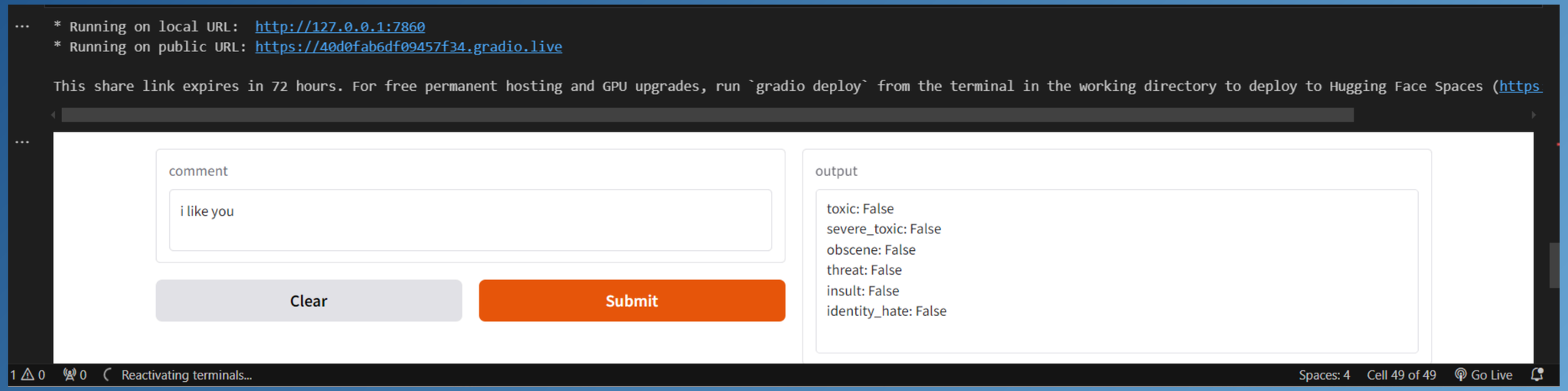
- BiLSTM (Bidirectional LSTM) reads text from both directions: left-to-right and right-to-left.
- This enables understanding of full sentence context, including negations and complex structures.
- Example: “I don’t hate you.” → BiLSTM detects the negation and understands it's non-toxic.
- Example: “You’re not stupid, just a little dumb.” → BiLSTM catches subtle insult others might miss.
- Why it matters: Cyberbullying often uses sarcasm, irony, or hidden threats that basic models can't catch.
- Compared to traditional RNNs or keyword models, BiLSTM greatly improves accuracy for toxicity detection.
- Summary: BiLSTM is ideal for cyberbullying detection due to its ability to deeply understand text in context.

MODELLING


User Input
User enters a comment for analysis
Text Preprocessing
Cleaning, tokenizing, and lowercasing the text
Text Vectorization
Converting text into numerical vectors
Embedding Layer
Transforms words into dense vectors
Bidirectional LSTM Model
Captures contextual meaning from both directions
Output Layer
Multi-label classification for toxicity, threat, insult, etc.
Gradio Interface
Displays result and flags comment in real-time

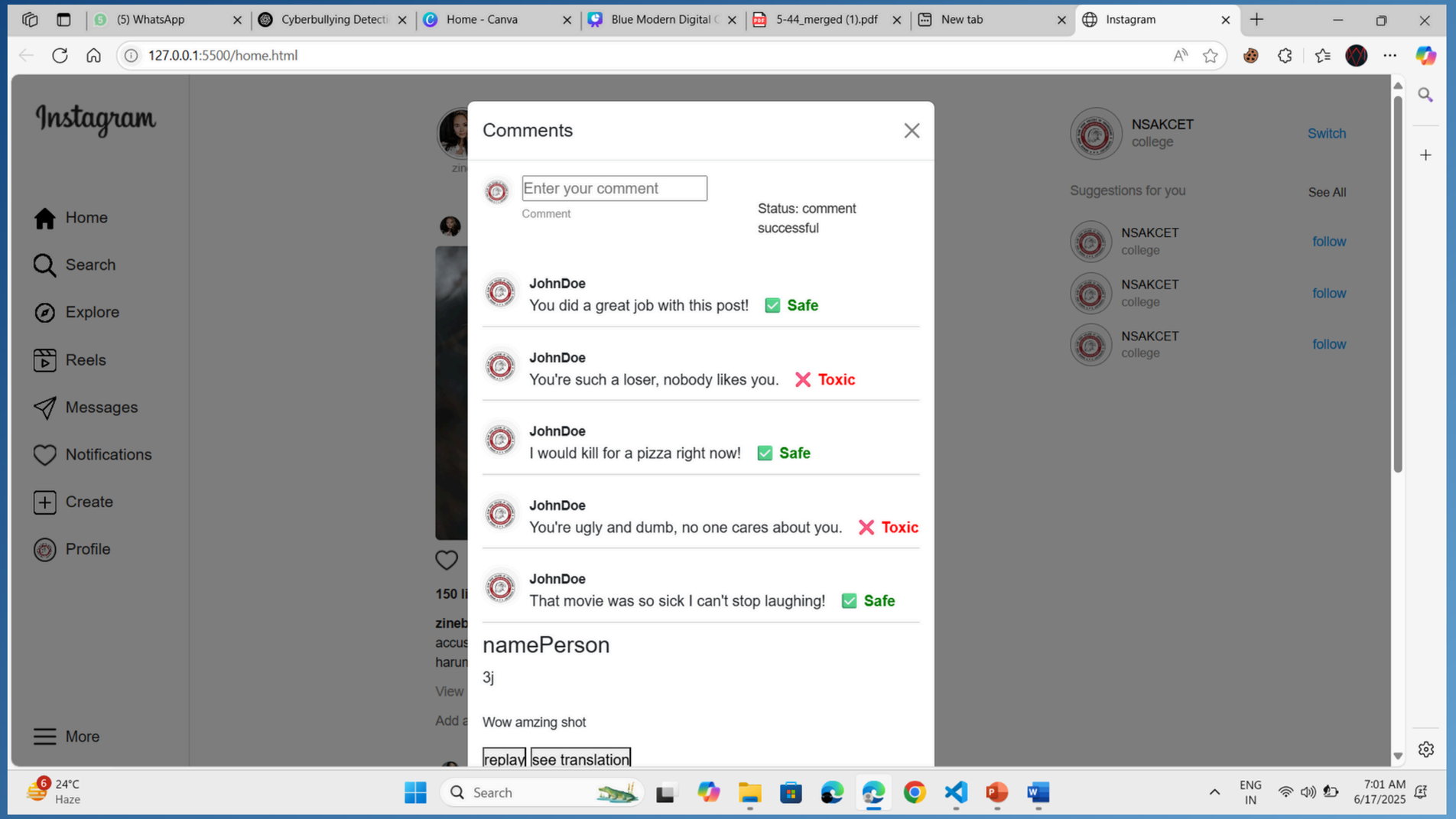
SCREENSHOTS





SCREENSHOTS

REAL TIME DETECTION



RESULT / OUTCOMES

- The model was able to detect and classify different types of harmful comments like toxic, threat, and insult.
- It gave good accuracy while testing and worked well even on tricky comments.
- I connected it with a Gradio interface and also created a basic Instagram-style UI, where users can enter comments like on a real post and instantly check if they are harmful.
- The system can be helpful in reducing cyberbullying by catching such comments early in platforms like social media.

CONCLUSION

This project helped me build a system that detects harmful comments like toxic, threat, or hate using a Bidirectional LSTM model. It was trained on real data and gave good results while testing.

I also added a Gradio interface and an Instagram-style UI to make it feel more practical and closer to how social media works. This project showed how deep learning and NLP can be used to reduce cyberbullying and make online spaces safer

FUTURE PERSPECTIVE

- Improve the model by using more advanced techniques like BERT or GPT for better accuracy.
- Add image or video comment analysis to detect bullying in visual content.
- Extend the system to work in different languages for wider reach.
- Integrate it with real social media platforms for live monitoring.
- Add an alert system to notify moderators when bullying is detected.

THANK YOU