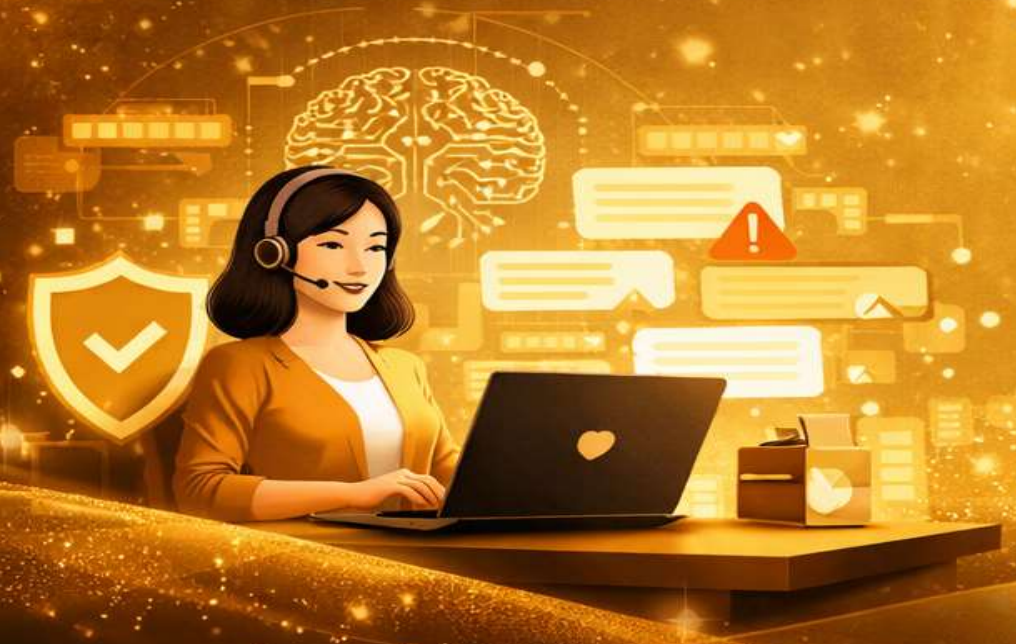
AI-Powered Comment Toxicity Detection System

Deep Learning Based Real-Time Moderation



Presented by: Abitha Jesuraj

Domain: NLP & Deep Learning

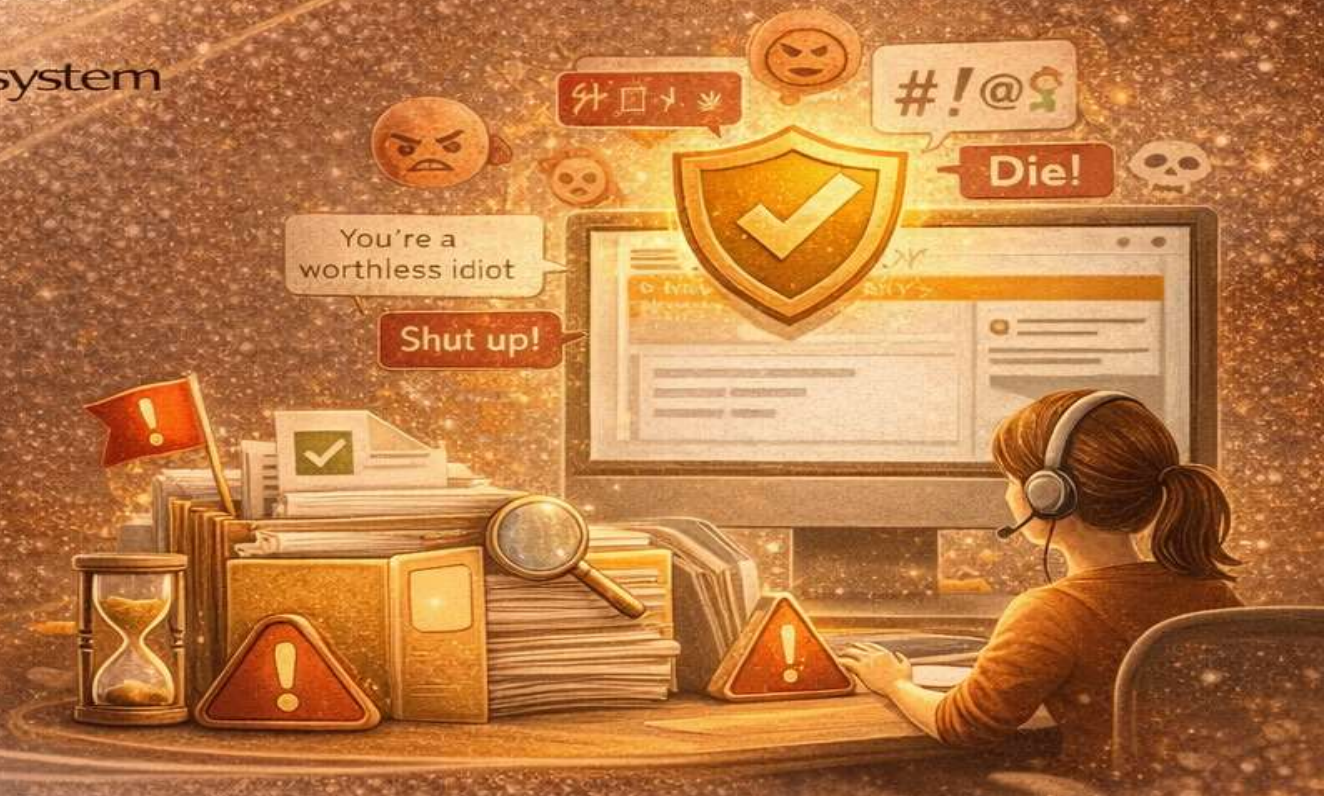
Project Overview

This project develops a Deep Learning-based Comment Toxicity Detection System to classify online comments as Toxic or Non-Toxic.

- Using NLP techniques and LSTM/CNN models, the system was trained and evaluated, with LSTM achieving the best performance (**96.13%** accuracy).
- The final model is deployed through a **Streamlit** dashboard that enables real-time prediction and bulk CSV moderation.

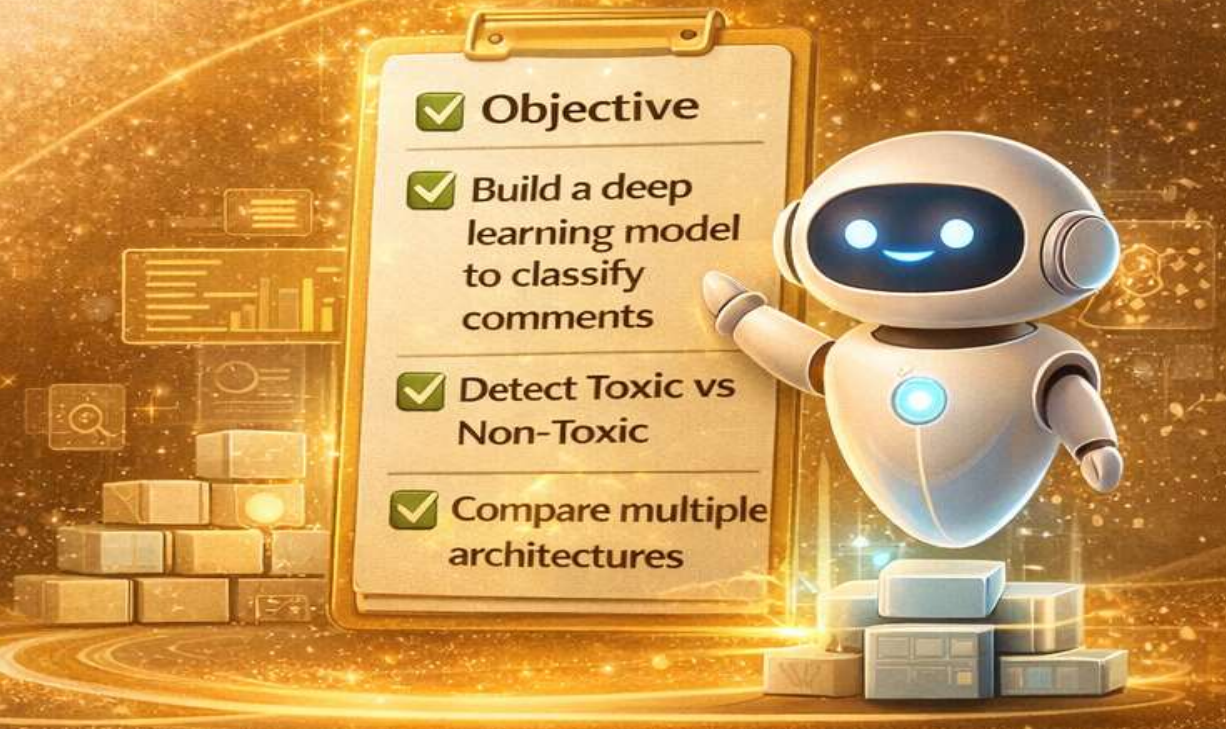
Problem Statement

- Online platforms face increasing toxic comments
- Includes harassment, abuse, hate speech
- Manual moderation is slow and inefficient
- Need automated real-time detection system



Objective

- Build a deep learning model to classify comments
- Detect Toxic vs Non-Toxic comments
- Compare multiple architectures
- Deploy using Streamlit dashboard



Dataset

Dataset: Jigsaw Toxic Comment Dataset

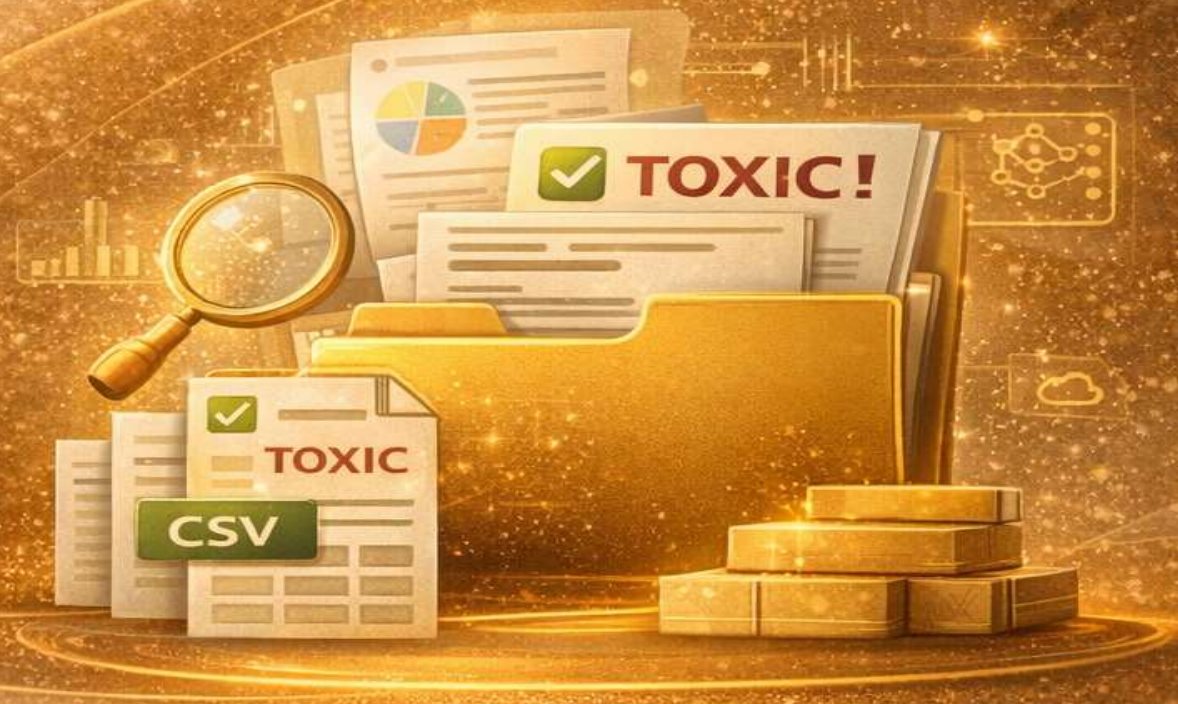
Original Labels:

- toxic
- severe_toxic
- obscene
- threat
- insult
- identity_hate

Binary Label Created:

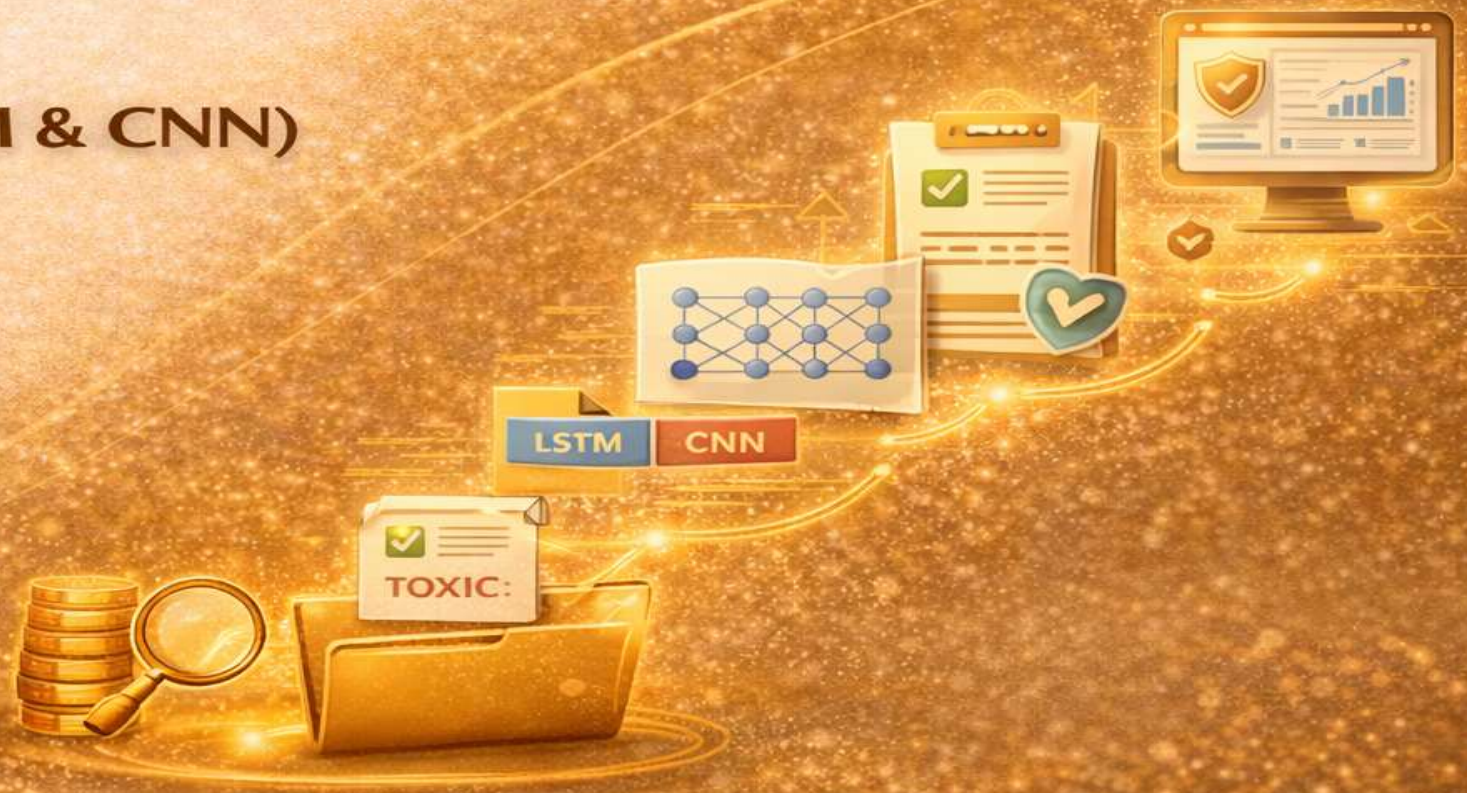
✅ Toxic (1) → If any label = 1

❌ Non-Toxic (0) → Otherwise



Project Workflow

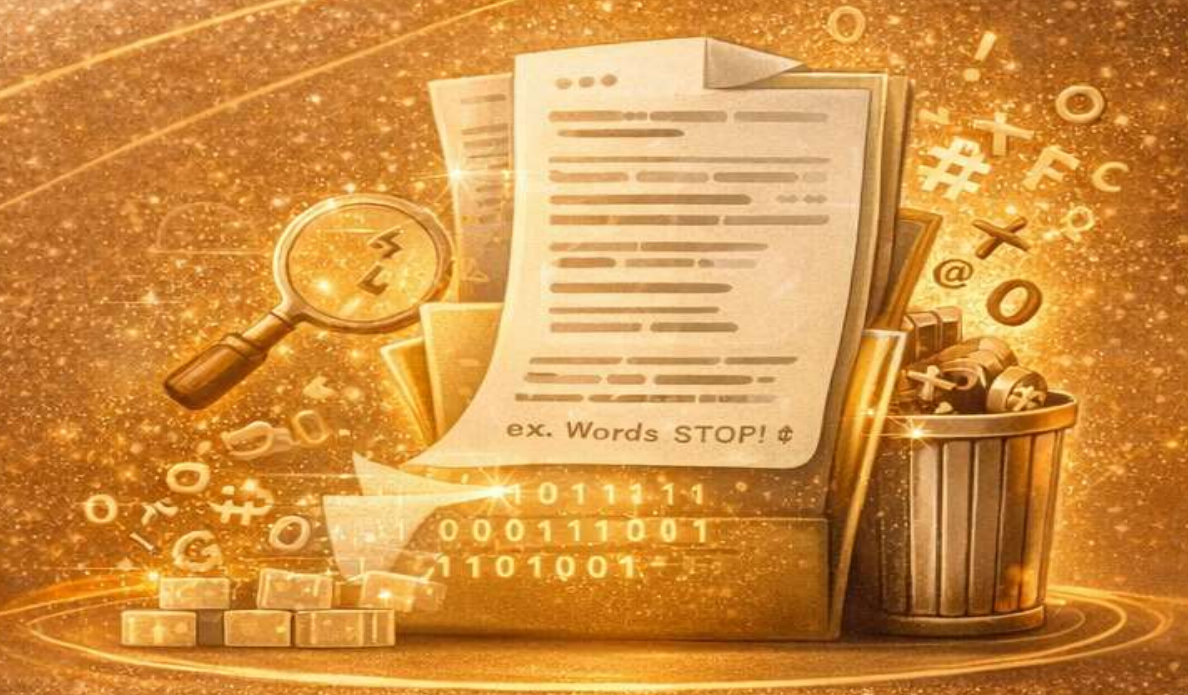
1. Data Exploration
2. Text Preprocessing
3. Model Training (LSTM & CNN)
4. Model Evaluation
5. Model Selection
6. Deployment



Text Preprocessing

- Lowercasing
- Removing special characters
- Stopword removal
- Tokenization
- Padding sequences
- Deployment

Purpose: Convert text into numerical format for model training



Model Development

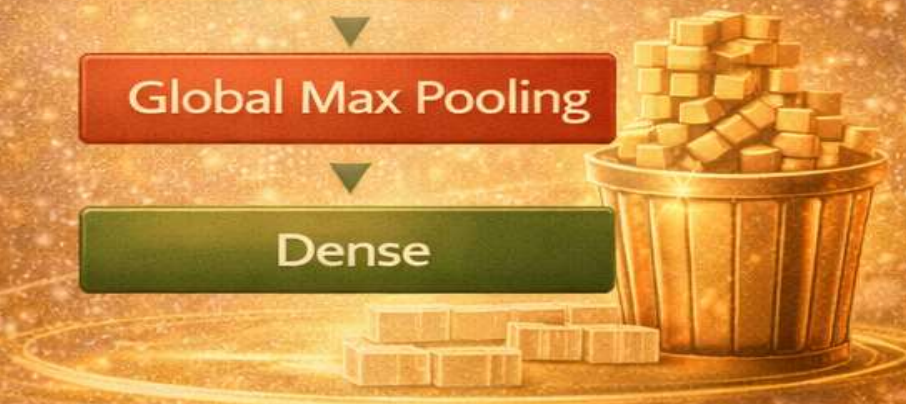
Two architectures implemented:

LSTM

- Embedding Layer
- LSTM Layer
- Dropout
- Dense (Sigmoid)

CNN

- Embedding Layer
- Conv1D
- Global Max Pooling
- Dense

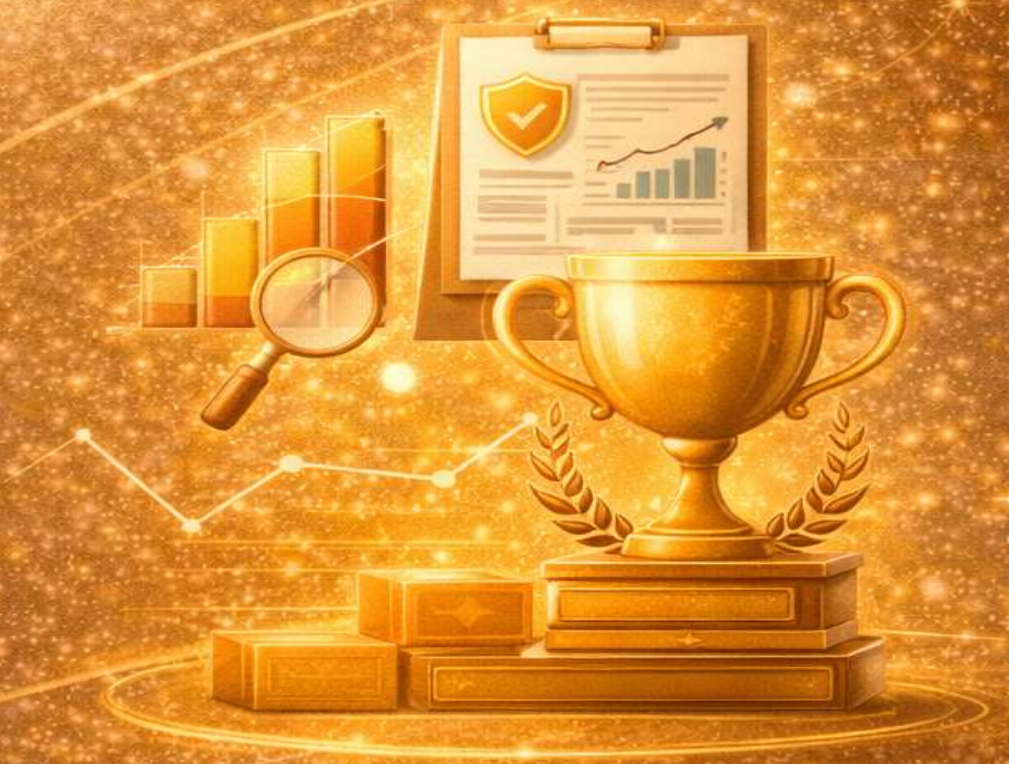


Purpose: Convert text into numerical format for model training

Model Performance

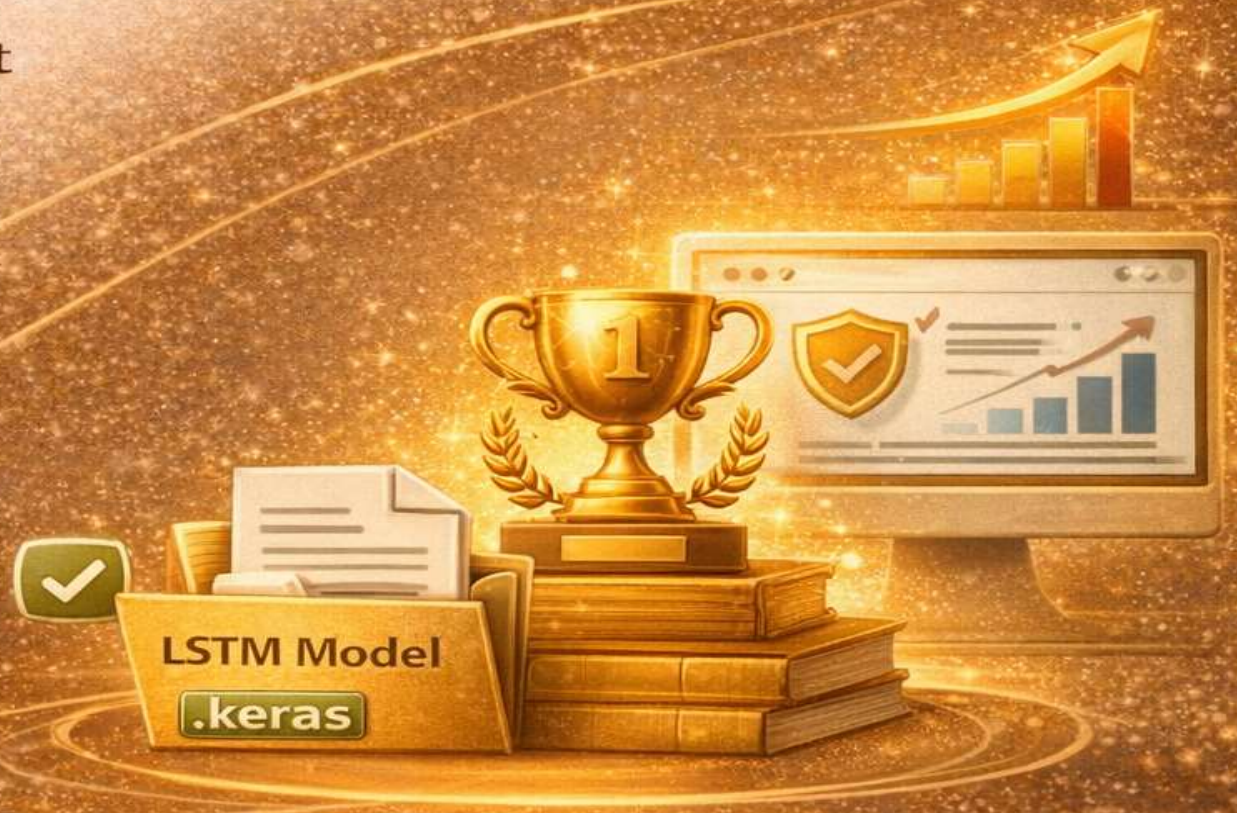
Model	Accuracy
LSTM	96.13%
CNN	95.71%

LSTM selected due to better contextual understanding.



Final Model Selection

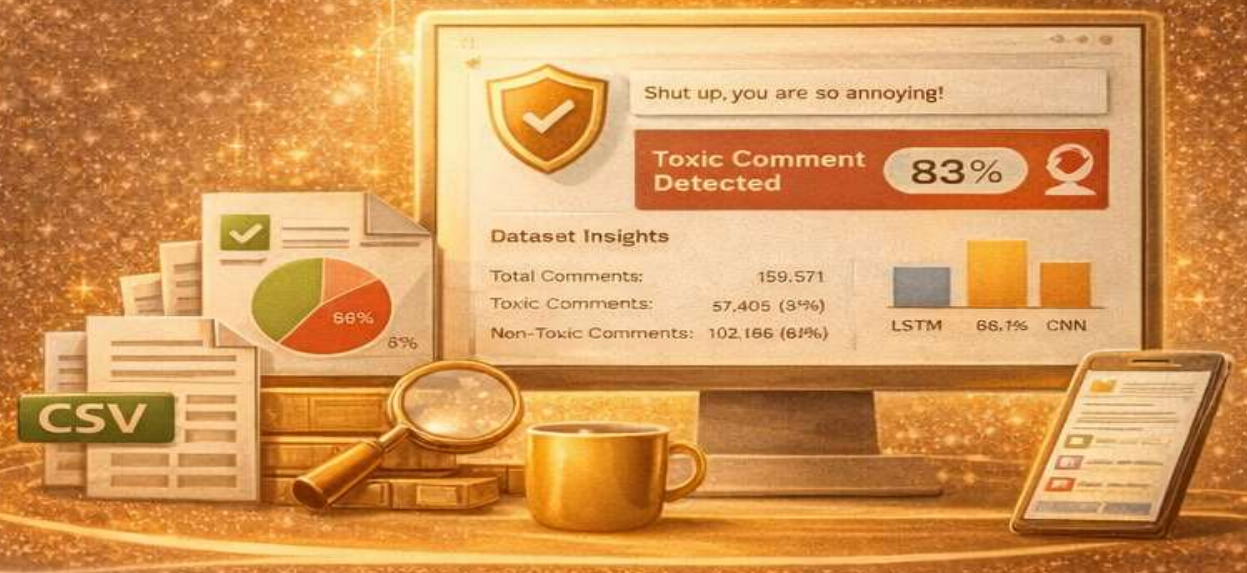
- LSTM achieved higher accuracy
- Better handling of sequence context
- Selected as deployment model
- Model saved in .keras format



Streamlit Deployment

Dashboard Features:

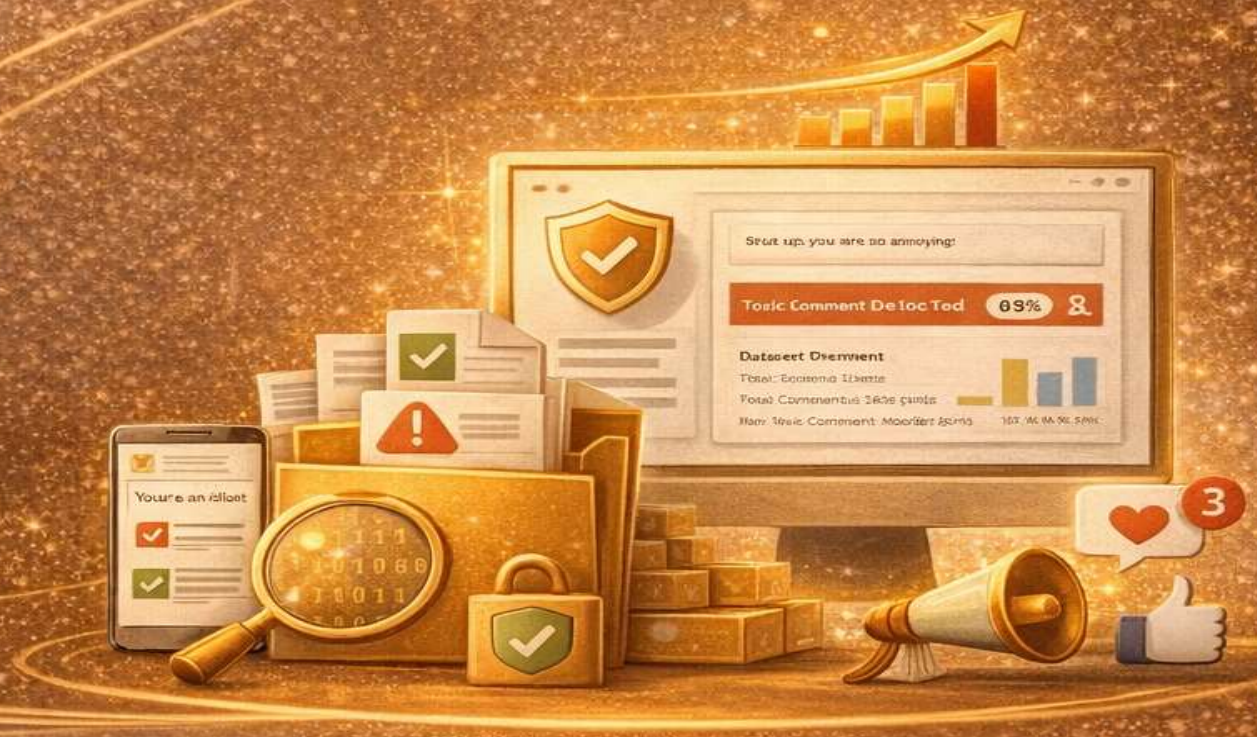
- Real-Time Prediction
- Bulk CSV Upload
- Dataset Insights
- Model Performance Visualization



Business Applications

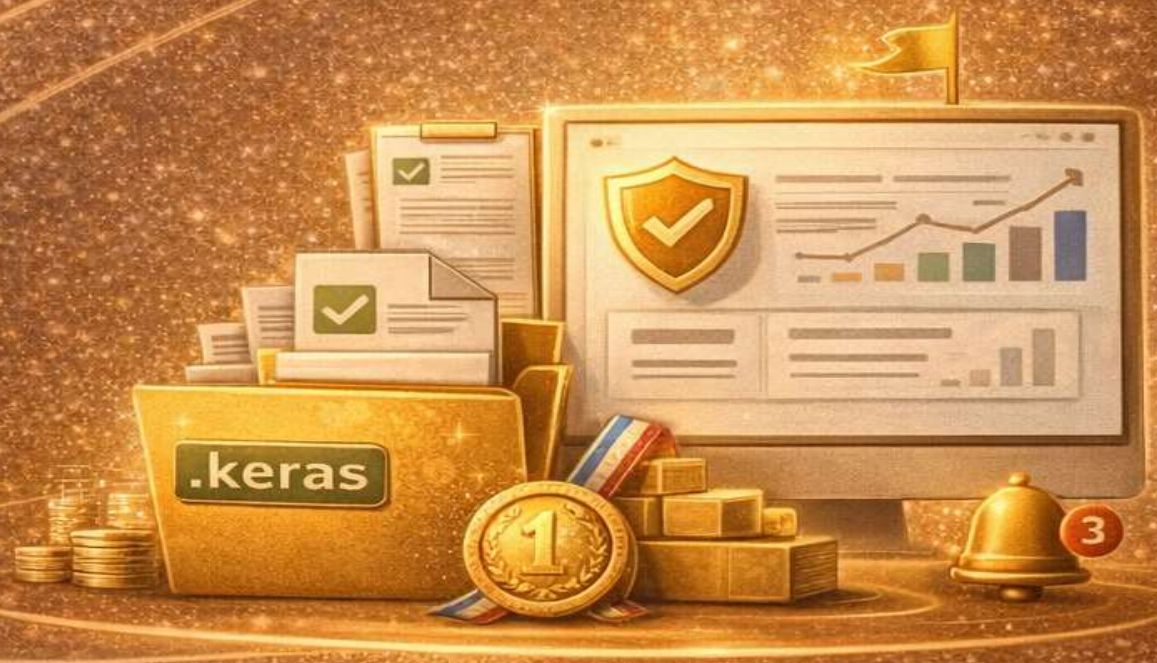
Dashboard Features:

- Social Media Moderation
- Community Forum Filtering
- Brand Safety Monitoring
- E-Learning Platforms
- Content Moderation Services



Conclusion

- Built end-to-end deep learning toxicity detection system
- Compared LSTM & CNN models
- Achieved 96% accuracy
- Successfully deployed using Streamlit



Thank You

Questions?