AI-Powered Online Harassment Detection System with Sentiment-Aware Decision Engine

Jyanesh Naidu
Department of Computer Science and Engineering (AIML)
Parul University, Vadodara, India
Email: alubillijyaneswarrao@gmail.com

***Abstract*—Online harassment poses a significant challenge to modern digital platforms due to its scale, anonymity, and evolving linguistic patterns. Manual moderation is inefficient, inconsistent, and difficult to scale. This paper presents an AI-powered online harassment detection system that integrates machine learning-based classification, sentiment analysis, and rule-based threat detection to enable real-time moderation. A sentiment-aware severity classification mechanism is introduced to reduce false positives caused by model uncertainty. The system is deployed on the cloud and provides an administrative dashboard for visualization and decision monitoring, closely reflecting real-world content moderation pipelines.**

***Index Terms*—Online Harassment Detection, Sentiment Analysis, Machine Learning, Natural Language Processing, Content Moderation, FastAPI**

I. Introduction

The exponential growth of online platforms has resulted in a corresponding increase in abusive and harassing content. Such behavior negatively impacts user well-being and platform trust. Traditional human-based moderation approaches are costly, slow, and prone to subjective bias. Consequently, automated AI-driven moderation systems have become essential to ensure platform safety at scale. This work focuses on designing an end-to-end, deployable harassment detection system that balances safety, fairness, and interpretability.

II. Problem Statement

Many existing automated moderation systems rely solely on probabilistic text classification models, leading to high false-positive rates. Neutral or positive content is often incorrectly flagged as harassment due to model uncertainty. The challenge is to accurately detect genuinely harmful content while avoiding unnecessary penalties on benign user behavior.

III. Proposed System

The proposed system follows a modular, multi-stage pipeline:

- Harassment probability estimation using supervised machine learning
- Sentiment polarity analysis to capture emotional intent
- Rule-based threat detection for explicit violent language
- Sentiment-aware severity classification
- Administrative decision engine with real-time visualization

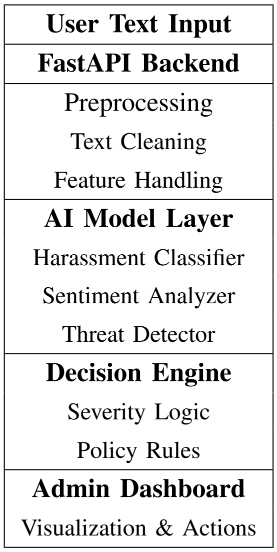

Fig. 1. System Architecture of the AI Harassment Detection Pipeline

A. System Architecture

IV. Methodology

A. Harassment Detection Model

A Logistic Regression classifier trained on labeled textual data is used to estimate the probability that an input message contains harassing content. The model outputs a continuous probability score to represent uncertainty.

B. Sentiment Analysis

The VADER sentiment analyzer from the NLTK library is employed to compute sentiment polarity scores ranging from -1 to +1. This allows the system to differentiate abusive intent from neutral or positive expressions.

C. Threat Detection

A rule-based NLP module detects explicit threats using keyword and pattern matching techniques. Threat detection is prioritized to ensure user safety.

D. Severity Classification

A sentiment-aware severity classification mechanism is introduced. Strong positive sentiment overrides uncertain harassment probabilities, effectively reducing false positives while maintaining moderation reliability.

E. Administrative Decision Engine

Based on severity and threat indicators, the decision engine produces one of four actions: *ALLOW*, *WARN*, *TEMP_BLOCK*, or *ESCALATE*.

V. Implementation

The backend service is implemented using FastAPI and deployed on the Render cloud platform. The administrative dashboard is built using Streamlit and communicates with the backend through REST APIs.

A. Technology Stack

- Python
- FastAPI
- Scikit-learn
- NLTK
- Streamlit
- Render Cloud

VI. Results and Evaluation

The system was evaluated using representative moderation scenarios.

TABLE I
Sample Moderation Results

Input Text	**Severity**	**Decision**
Happy	LOW	ALLOW
You are stupid	MEDIUM	WARN
I will kill you	HIGH	ESCALATE

The inclusion of sentiment-aware logic significantly reduced false positives compared to probability-only classification approaches.

VII. Challenges Faced

Several challenges were encountered during development:

- GitHub authentication and deployment configuration issues
- Missing serialized model files during cloud deployment
- NLTK resource loading errors in production
- False positives caused by model uncertainty

These challenges were resolved through systematic debugging, path correction, and decision logic refinement.

VIII. Conclusion

This work demonstrates a practical and scalable approach to online harassment detection by integrating machine learning with sentiment-aware decision logic. The system is deployable, interpretable, and aligned with real-world moderation practices, making it suitable for large-scale online platforms.

IX. Future Work

Future enhancements include:

- Transformer-based NLP models such as BERT
- Multilingual harassment detection
- Image and video-based moderation
- User behavior and temporal analytics