# PROJECT PROPOSAL

# AI-Based Spam and Abuse Detection System for Online Comment Platforms

**By:**

**Ricus Yeshua M. Alzona**

## Introduction

The rapid expansion of the internet and social media platforms has transformed the way people communicate, express opinions, and engage in public discourse. Online comment platforms, such as those found on social networking sites, blogs, news websites, and discussion forums, allow users to interact freely and share ideas in real time. While this openness promotes democratic participation and information exchange, it also creates opportunities for misuse, including the spread of spam, harassment, hate speech, and other forms of abusive behavior.

Spam and abusive comments not only degrade the quality of online discussions but also discourage meaningful participation, particularly among vulnerable users. In many cases, these comments can escalate into cyberbullying, misinformation, and toxic online environments. As comment volumes increase, especially on popular platforms, manual moderation becomes inefficient, costly, and inconsistent. This situation highlights the urgent need for an intelligent and automated approach to content moderation.

This project proposes the development of an AI-Based Spam and Abuse Detection System designed to automatically analyze user comments, identify harmful content, and support moderators in enforcing community guidelines. By integrating artificial intelligence into the moderation process, the system aims to enhance accuracy, efficiency, and scalability while maintaining human oversight.

## Background of the Study

Online platforms today face increasing pressure to moderate user-generated content effectively. Traditional moderation approaches rely heavily on human moderators or basic rule-based filters that detect prohibited keywords. While these methods provide some level of control, they are often insufficient when dealing with context-dependent language, sarcasm, coded insults, or rapidly evolving spam tactics.

Manual moderation is labor-intensive and prone to fatigue and bias, especially when moderators are exposed to large volumes of offensive content. Additionally, delays in moderation may allow harmful comments to remain visible for extended periods, causing emotional distress to users and reputational damage to platforms. Rule-based systems, on the other hand, lack adaptability and can generate false positives or miss subtle forms of abuse. Advances in artificial intelligence, particularly in natural language processing (NLP) and machine learning, have made it possible to analyze text more intelligently. AI systems can learn patterns of abusive language, spam behavior, and contextual meaning from large datasets. These technologies enable automated moderation systems to detect harmful content with greater accuracy and consistency.

This study is motivated by the need to apply AI techniques to online comment moderation in order to create a safer, more inclusive digital environment. The proposed system focuses on combining automated AI analysis with human moderation to achieve balanced and effective content control.

## Statement of Objectives

### General Problem

Online comment platforms lack a reliable, scalable, and intelligent system for automatically detecting and managing spam and abusive content, resulting in poor content quality, unsafe user interactions, and inefficient moderation processes.

### Specific Problems

- The increasing volume of online comments makes manual moderation impractical and inefficient.
- Spam comments clutter discussion threads and reduce the credibility of online platforms.
- Abusive language, harassment, and hate speech negatively affect user well-being and participation.
- Existing keyword-based filtering systems fail to capture contextual and nuanced forms of abuse.
- Moderators often lack centralized tools to track flagged content and moderation decisions.
- Inconsistent moderation decisions lead to unfair enforcement of platform rules.

## General Objective

To design and propose an AI-Based Spam and Abuse Detection System that automatically analyzes online comments and assists moderators in maintaining a safe, respectful, and high-quality online discussion environment.

## Specific Objectives

- To develop a system that automatically analyzes user comments using AI-based text processing techniques.
- To classify comments into approved, flagged, or blocked categories based on spam and abuse scores.
- To provide real-time feedback to users when their comments violate content guidelines.
- To design a moderator dashboard that allows efficient review and management of flagged comments.
- To maintain logs of AI decisions and moderator actions for transparency and accountability.
- To improve moderation efficiency while reducing the workload of human moderators.

## Scope and Limitation

### Scope of the Study

- The proposed system will focus on the moderation of text-based comments submitted to an online platform. The system will:
- Allow users to register, log in, and post comments.
- Automatically analyze comments upon submission using AI.
- Generate spam and abuse confidence scores for each comment.
- Automatically approve, flag, or block comments based on predefined thresholds.
- Provide moderators with a dashboard to review flagged comments.
- Allow moderators to approve, delete, or take action against abusive users.
- Store data related to users, comments, AI analysis results, and moderation actions.
- Generate basic reports for monitoring system performance and moderation trends.
- Limitations of the Study
- Despite its capabilities, the system has several limitations:
- The AI model may produce false positives or false negatives, especially with ambiguous language.
- The system does not analyze multimedia content such as images, videos, or audio.
- The system’s accuracy depends on the quality and diversity of training data.
- The system does not fully replace human moderators and still requires human judgment.
- Language support may be limited to selected languages.

## The Proposed System

The proposed AI-Based Spam and Abuse Detection System serves as an intelligent moderation layer within an online comment platform. It automatically evaluates comments before or shortly after they are posted, reducing exposure to harmful content. The system combines automated AI decision-making with human oversight to ensure balanced and fair moderation.

The system is designed to be modular, scalable, and adaptable to different types of online platforms, including forums, blogs, and social media websites.

### System Overview

The system is composed of three main modules that work together to achieve effective content moderation:

#### User Interface Module

This module allows users to interact with the platform by creating accounts, logging in, and posting comments. Users receive feedback when their comments are flagged or blocked, promoting awareness of acceptable online behavior.

#### AI Analysis Module

The AI Analysis Module processes submitted comments using natural language processing techniques. It evaluates text patterns, keywords, sentiment, and contextual indicators to determine whether a comment contains spam or abusive content. Based on calculated confidence scores, the system automatically classifies comments and forwards flagged content to moderators for review.

### Moderator and Administrator Module

This module provides moderators with tools to efficiently manage flagged comments. Moderators can review AI decisions, approve or delete comments, and take action against abusive users. Administrators can access reports, monitor moderation trends, and evaluate system performance.