Product Requirements Document: Advanced Admin & Moderation Platform

1. Introduction

This PRD outlines the requirements for an enhanced admin and moderation platform. Building upon foundational moderation tools and content management features, this document details an expanded suite of tools designed to improve efficiency, automate workflows, and proactively ensure a safe and positive user experience.

2. Target Audience

This platform is for internal employees with various moderation and administrative roles, including:

- **Moderators**
- **Trust & Safety Team**
- **Community Managers**
- **Administrators**

3. Problem Statement

Manual moderation is inefficient, reactive, and doesn't scale with user growth. Proactively identifying and addressing harmful behavior requires sophisticated tools, automation, and a consolidated view of user activity. Without these, our platform is vulnerable to spam, abuse, and inconsistent policy enforcement.

4. Solution Overview

The solution is a centralized platform that integrates **advanced moderation tools**, **proactive automation**, and **comprehensive analytics**. It will empower moderators to work more efficiently, enable Trust & Safety teams to identify and neutralize threats, and provide administrators with a complete picture of platform health.

5. Functional Requirements

5.1. Automated Moderation & Proactive Tools

- **5.1.1. Automated Content Flagging:** The system must use machine learning and NLP to automatically flag content for:
    - **Text Violations:** Hate speech, harassment, spam, and profanity, including in disguised forms (e.g., "h4t3").
    - **Media Violations:** Violence, nudity, and explicit content in images and videos.
- **5.1.2. User Reputation System:** A user's trust score will be calculated based on their positive or negative actions. High-score users' reports will be prioritized. Content from low-score users may be automatically placed in a pre-moderation queue.
- **5.1.3. Ban Evasion Detection:** The platform will use **IP addresses and device**

fingerprints to identify users attempting to create new accounts after a ban, automatically flagging them for review.
- **5.1.4. Shadowbanning:** The ability to make a user's content visible only to them, effectively hiding it from the community without their knowledge. This is a key tool for neutralizing spammers.

5.2. Content Moderation & User Management

- **5.2.1. Centralized Moderation Queue:** A single queue will display all user-reported and system-flagged content, prioritized by severity.
- **5.2.2. Comprehensive User Profile:** An enhanced view of each user's profile showing their **complete action history**, including all posts, comments, and private notes from previous moderation actions.
- **5.2.3. Moderation Actions:**
    - **Muting & Restrictions:** Admins can temporarily mute a user's ability to interact or restrict their posting privileges.
    - **Permanent & Temporary Ban:** Banning a user must require a reason, which is logged for transparency.
- **5.2.4. User Appeal System:** Users can appeal bans or content removals through a dedicated, on-site system. The appeal will be routed to the moderation team for review.

5.3. Workflow & CMS Tools

- **5.3.1. Content Management System (CMS):** A WYSIWYG editor allows non-technical staff to update public-facing documents like the Terms of Service and Privacy Policy.
- **5.3.2. Version Control:** The CMS must track all changes, including the author, timestamp, and a change log. It should also allow for scheduled publishing.
- **5.3.3. Moderator Collaboration:** The platform will allow moderators to leave private notes on user profiles or specific content for team context and historical reference.
- **5.3.4. Saved Responses:** A library of templated messages for common user inquiries or moderation actions to ensure consistent communication.

5.4. Reporting & Auditing

- **5.4.1. Detailed Activity Log:** A comprehensive log of every moderation action, including the moderator's ID, the action taken, and the timestamp. This log must be easily searchable.
- **5.4.2. Analytics Dashboard:** The dashboard will display key metrics and trends, such as daily content reports, ban rates, and a breakdown of violations by category.
- **5.4.3. Transparency Reports:** The system can generate public-facing reports on moderation activity to build trust with the community.

6. Technical Requirements

6.1. Front-End (UI/UX)

- **6.1.1. Intuitive Dashboard:** A clean and easy-to-navigate interface that prioritizes common tasks.

- **6.1.2. Role-Based UI:** The UI will adjust based on the employee's role, showing only the tools and data they are permitted to see.

6.2. Back-End

- **6.2.1. API Endpoints:** New endpoints are required for all advanced features, including POST /api/moderation/user/{id}/shadowban and POST /api/moderation/content/{id}/report.
- **6.2.2. Database Schema:** New tables are needed to support the extended functionality, including user_reputations, moderation_notes, and a more robust logging system.
- **6.2.3. ML Integration:** A microservice or external API will handle the automated content analysis and flagging.
- **6.2.4. Access Control: Role-Based Access Control (RBAC)** will be strictly enforced at the API level to ensure secure access to sensitive data and tools.

7. Success Metrics

- **Moderator Efficiency:** A significant decrease in the average time to resolve flagged content.
- **Community Health:** A measurable reduction in user-reported violations.
- **Team Satisfaction:** Positive feedback from the moderation team on the platform's usability and effectiveness.
- **Automation Effectiveness:** The percentage of violating content successfully flagged by the automated system.