Production MLOps for Spam Detection: Real-Time Multi-Modal AI at Billion-User Scale

By Prabhakar Singh

Meta

Conf42.com MLOps 2025

Today's Agenda

01

Addressing the Billion-User Scale

Understanding the unique requirements and constraints of spam detection at a massive scale.

02

End-to-End MLOps Architecture

Designing robust pipelines for real-time, multi-modal detection.

03

Intelligent Data Strategies

Leveraging advanced sampling, labeling, and privacy-preserving techniques.

04

High-Stakes Production Engineering

Scaling, monitoring, and maintaining critical AI systems.

05

Charting Future Directions

Exploring emerging threats and innovative technologies in spam detection.

This session offers actionable technical strategies and architectural patterns for ML engineers, platform architects, and MLOps practitioners building high-scale content moderation and fraud detection systems.

The Scale Challenge

Processing Requirements

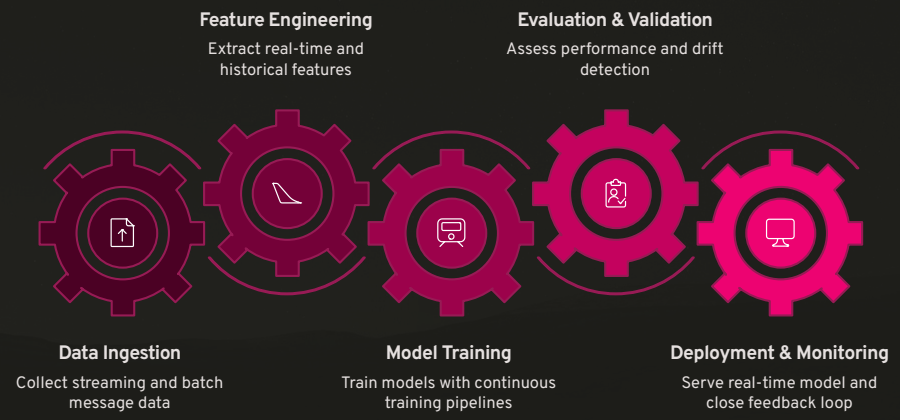
- Processing billions of content pieces daily
- Handling millions of evaluations per second
- Meeting sub-100ms response time SLAs
- Ensuring 24/7 availability with zero-downtime deployments

Technical Constraints

- Detecting spam across multi-modal content (text, images, video)
- Supporting multilingual detection capabilities
- Achieving 99%+ accuracy
- Adapting rapidly to evolving attack vectors
- Adhering to strict privacy compliance (e.g., GDPR, CCPA)

These demanding requirements present a unique MLOps challenge, where even a 0.1% false positive rate can adversely affect millions of legitimate users daily.

End-to-End MLOps Architecture



Modern spam detection systems require sophisticated MLOps pipelines that support continuous improvement while maintaining 24/7 reliability.

Multi-Modal Feature Engineering

Text Features

- Transformer-based embeddings (e.g., BERT variants)
- N-gram frequency analysis
- Named entity recognition
- Language-specific tokenization
- Sentiment and intent classification

Image & Video Features

- CNN-based embeddings (e.g., EfficientNet, ResNet)
- Object detection and classification
- OCR for embedded text
- Frame sequence analysis
- Perceptual hashing for near-duplicate detection

Behavioral Features

Account age, posting patterns, social graph connections, engagement metrics, device fingerprinting

Contextual Features

Time of day, geographic distribution, trending topics, platform-specific signals, cross-post correlation

Effective feature engineering combines signals across multiple modalities to detect sophisticated attacks that might appear benign in any single dimension.

Model Architecture: Ensemble Approach

Fast Pre-filters

Lightweight models (e.g., GBDTs, simple neural networks) for rapid, high-recall assessment.

Specialist Models

Domain-specific deep learning models tailored for text, image, video, and user behavior data.

Meta-learner

A final decision layer that weighs specialist model outputs based on confidence and historical performance.

Multi-modal Fusion

Cross-attention mechanisms to correlate and integrate signals across diverse content types.

This hierarchical approach ensures both computational efficiency and high accuracy, with specialist models dynamically activated based on content and context.

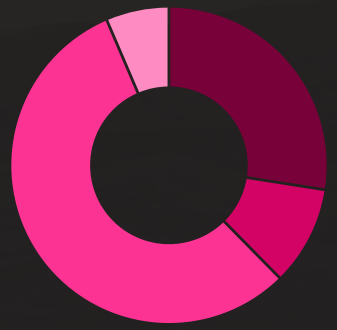
Distributed Serving Architecture

Core Components

- **Content-aware load balancing:** Efficiently routing requests.
- **Horizontal auto-scaling:** Dynamically adjusting resources.
- **Low-latency feature store:** Centralized, performant feature storage.
- **GPU/TPU acceleration:** Harnessing specialized hardware for inference.

Performance Optimization

- **Model quantization:** Reducing model size and accelerating inference.
- **Batching strategies:** Processing multiple requests simultaneously.
- **Request prioritization:** Ensuring critical tasks are handled first.
- **Circuit breakers:** Preventing cascading failures during overload.



Feature Extraction | Pre-filter Models | Specialist Models | Ensemble Decision

Our architecture achieves an 80ms end-to-end latency, meeting the target SLA.

Smart Sampling & Labelling Strategies

Uncertainty Sampling

Direct human review towards content where model confidence is low.

Diversity Sampling

Ensure diverse data representation to mitigate bias in training sets.

Adversarial Sampling

Identify examples that exploit model weaknesses to enhance robustness.

Time-Sensitive Sampling

Dynamically adjust sampling based on temporal patterns and emerging attack vectors.

These strategies collectively optimize human review processes, significantly reducing labelling efforts and enhancing overall model performance.

Coordinated Attack Detection

Real-time Streaming Analysis

Process millions of actions per second to identify temporal patterns and statistical anomalies indicative of coordinated behavior.

Graph Neural Networks

Leverage Graph Neural Networks to model user-content interactions as dynamic graphs, detecting suspicious clusters and propagation patterns characteristic of coordinated campaigns.

Proactive Countermeasures

Implement targeted throttling, CAPTCHA challenges, and heightened scrutiny based on real-time threat assessments.

Sophisticated spam attacks often involve thousands of accounts operating in coordination. While individual actions may appear benign, collective analysis across the platform reveals distinct, actionable patterns.

Privacy-Preserving ML Techniques

Federated Learning

Trains models across decentralized data silos, eliminating the need to centralize sensitive user information:

- Secure aggregation protocols
- Differential privacy guarantees
- On-device inference capabilities

Encryption Technologies

Processes data securely, ensuring privacy throughout the lifecycle:

- Homomorphic encryption for secure inference
- Secure multi-party computation
- Zero-knowledge proofs for compliance
- Private set intersection for secure lookups

These advanced techniques ensure compliance with stringent privacy regulations like GDPR and CCPA, critically balancing data protection with robust detection efficacy.

Continuous Training & Deployment

Model Training

Daily retraining with new data, utilizing Bayesian optimization for hyperparameter tuning.

Shadow Deployment

New models run in parallel with production traffic, evaluating performance without impacting user experience.

Evaluation

Comprehensive testing against golden datasets, adversarial examples, and fairness benchmarks ensures model integrity.

Progressive Rollout

Gradual rollout via canary testing, with automated analysis of performance metrics and controlled traffic shifting.

Production Monitoring & Observability

Model Performance Monitoring

- Precision and recall drift detection
- Slice-based evaluation across diverse demographics
- Confusion matrix evolution over time

System Performance Monitoring

- Request latency distributions (p50, p95, p99)
- Queue depths and backpressure metrics
- Resource utilization (CPU, GPU, memory)

Data Drift Monitoring

- Feature distribution shifts
- Data quality metrics and schema validation
- Statistical tests for significance of changes

Business Impact Monitoring

- Impact of false positives on user engagement
- Impact of false negatives on platform health
- Tracking content deletion appeals and their success rates

Comprehensive monitoring allows for early detection of both technical issues and emerging spam tactics, enabling proactive response before widespread impact.

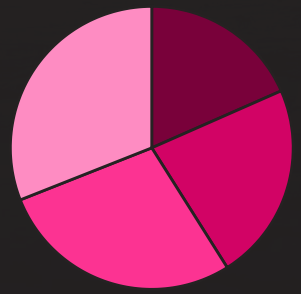
Anomaly Detection & Incident Response

Automated Anomaly Detection

- Statistical outlier detection on key metrics
- Multivariate time series analysis
- Prediction-based deviation detection
- Seasonality-aware baseline comparison
- Internal model consistency checks

Structured Incident Response

- Severity-based escalation paths
- Automated model rollbacks
- Shadow mode for investigation
- Configurable safety thresholds
- Model-specific debugging tooling
- Post-mortem analysis framework



Q1 2024 Q2 2024 Q3 2024 Q4 2024

Emerging Challenges & Future Directions

Synthetic Content Detection

Identify synthetic text, images, and deepfakes generated by accessible AI that bypass traditional filters.

Cross-Platform Coordination

Coordinate detection systems across platforms to share signals while preserving privacy.

Adversarial Robustness

Strengthen defenses against sophisticated adversarial attacks exploiting model weaknesses.

Explainable AI Integration

Provide clear explanations for enforcement actions, supporting appeals and transparency.

The arms race between spam detection and attackers demands continuous innovation.

Key Takeaways

1 Multi-Modal Integration is Essential

Effective spam detection demands integrated signal processing across all modalities—text, images, video, and user behavior—to combat sophisticated, multi-faceted attacks.

2 MLOps at Scale Requires Specialized Architecture

For billion-user platforms, effective MLOps necessitates purpose-built infrastructure, including intelligent serving tiers, distributed processing, and highly automated deployment pipelines.

3 Privacy and Performance Can Coexist

Achieving high-accuracy detection while upholding stringent privacy protections is possible through advanced techniques like federated learning and homomorphic encryption.

4 Continuous Adaptation is Non-Negotiable

Given the daily evolution of spam tactics, platforms must adopt continuous adaptation through comprehensive monitoring, anomaly detection, and rapid deployment capabilities to maintain effectiveness.

These principles are critical for building robust, scalable, and privacy-preserving AI systems capable of combating evolving threats at global scale.

Thank You !