COLLECTION OF PAPERS
NEW ECONOMY
Volume 4, No. 1, 2026

DESIGN AND IMPLEMENTATION OF A MULTI-MODEL AI-BASED CONTENT MODERATION SYSTEM IN A PRODUCTION ENVIRONMENT

Maksym Nenashev[1*]

Received 01.03.2026. | Send to review 18.03.2026. | Accepted 19.08.2026.

Original Article

[1] Wyższa Szkoła Biznesu - National Louis University (WSB-NLU)

***Corresponding Author:**
Maksym Nenashev

Email:
manenashev@wsb-nlu.edu.pl

JEL Classification:
C45, L86, O33

Doi: 10.61432/CPNE0401105n

UDK: 004.896:351.824.11(4)

ABSTRACT

This paper presents the design and implementation of an AI-based multimodal content moderation system deployed in a real-world production environment. The study addresses the scalability and effectiveness challenges associated with moderating large volumes of user-generated text and image content. The proposed system integrates natural language processing, convolutional neural networks, and embedding-based facial recognition within an ensemble of eight parallel models to improve classification accuracy, robustness, and redundancy.

Experimental results demonstrate strong performance, with text-classification accuracy exceeding 90% and image-detection accuracy reaching 92%, while maintaining end-to-end latency below one second. These findings confirm the practical value of multimodel architectures for large-scale digital platforms and highlight their potential to improve platform safety, operational efficiency, and broader socio-economic outcomes.

Keywords: *Artificial Intelligence, Content Moderation, Multi-model Architecture, NLP, Computer Vision, Face Recognition, MLOps, Risk Control*

1. INTRODUCTION

The rapid expansion of global online platforms and social media ecosystems has generated an unprecedented volume of user-generated content. This growth has intensified the challenges associated with content moderation, particularly the detection of inappropriate, violent, harmful, or sensitive material. Manual moderation is no longer sufficient because it cannot provide the scalability, consistency, and response speed required by high-traffic digital platforms (Vankov & Lukarev, 2025). Consequently, automated AI-driven moderation systems have become essential for maintaining platform safety and operational efficiency (Nassanbekova et al., 2026). This study presents the design and deployment of a multimodel content moderation system in a live production environment.

Although single-model architectures have been widely examined, recent advances in machine learning operations (MLOps) and ensemble learning underscore the importance of multimodal approaches in production settings (Goodfellow et al., 2016). Reliance on isolated classifiers may result in high false-positive rates, limited robustness, and increased vulnerability to concept drift in continuously evolving user-generated content. Integrating natural language processing, computer vision (He et al., 2016), and ensemble-learn-

ing techniques (Dietterich, 2000) can improve classification reliability, fault tolerance, and system resilience.

This study contributes to this practical research domain by presenting a scalable MLOps pipeline that connects theoretical multimodel architectures with high-load production deployment (Mäkinen et al., 2021). To address localized concept drift and multilingual variation, the proposed architecture also supports dedicated contextual language models tailored to specific linguistic groups. This design enables more precise and context-sensitive moderation across diverse user populations.

2.PROBLEM FORMULATION

Modern digital ecosystems generate vast volumes of real-time data, requiring content moderation frameworks that provide:

- high scalability to accommodate fluctuating traffic volumes;
- rapid processing to support real-time user interactions;
- high classification accuracy to preserve platform integrity; and
- fault tolerance and robustness against classification errors.

The diversity of user-generated content, ranging from text and images to complex multimodal inputs, further increases the technical and operational complexity of the moderation process (Bouzaidi et al., 2026).

3. SYSTEM ARCHITECTURE

The technical architecture of the proposed moderation framework is designed around the principles of distributed processing and model redundancy. Unlike traditional monolithic systems that process data sequentially, this architecture separates user interactions from heavy analytical tasks. The primary objective of this design is to ensure transparent, reliable, and predictable system outcomes while maintaining continuous platform availability even during unexpected operational loads.

By isolating individual system components, the framework guarantees that a peak load or a temporary slowdown in one module - such as the image analysis or biometric processing units - does not degrade the overall user experience. This high degree of modularity allows the platform to balance computational resources dynamically, ensuring that critical safety checks are completed efficiently without impacting the core business operations of the digital ecosystem.

3.1. DATA PROCESSING LAYER

The underlying infrastructure supports the production environment of FindWay, a specialized social platform and international search network. As a fully functional social platform, it processes a diverse and continuous stream of user-generated content, including user profiles, media attachments, public comment sections, and direct messaging. Managing these real-time collaborative interactions requires a robust data management strategy capable of handling complex data packets without causing backend latency or service interruptions.

To achieve this stability, the system employs an asynchronous data-processing pipeline based on distributed queues. When a user uploads a post or sends a message, the content is temporarily placed in a secure queue and immediately released to the front-end interface, while heavy AI classification models process the payload in the background. This structural approach eliminates server overloads, reduces continuous infrastructure maintenance costs, and ensures a seamless, real-time user experience across all interactive features of the network.

3.2. COMPUTER VISION MODULE

For image analysis and the detection of Not Safe for Work (NSFW) content, the system employs the EfficientNet-B0 architecture (Tan & Le, 2019). This model was selected because it provides an effective balance between computational efficiency and classification accuracy, making it suitable for high-volume production environments.

3.3. NATURAL LANGUAGE PROCESSING MODULE

The natural language processing module is designed to detect toxic content and perform semantic classification across diverse linguistic environments. To ensure high precision, the module utilizes an ensemble of eight specialized models (versions v3 through v6.1). Each model is calibrated for specific linguistic groups and contextual settings, allowing the system to maintain consistent moderation accuracy despite regional dialectal variations. This tiered approach provides complementary and overlapping coverage, enabling the platform to adapt to evolving user-generated content. Future development will focus on integrating additional contextual models to further refine this classification capability and expand support for emerging languages.

3.4. MULTI-MODEL ENSEMBLE

The system's final decision layer consists of eight AI models operating in parallel. This ensemble architecture provides three principal advantages:

- Redundancy: Overlapping model functions reduce dependence on individual classifiers and prevent single points of failure.
- Individual calibration: Each model is independently tuned to reduce false-positive rates and improve risk-sensitive classification.
- Aggregated decision-making: Model outputs are combined to generate the final moderation decision, thereby improving robustness, consistency, and overall system stability.

Figure 1. High-level architecture of the multi-model moderation system

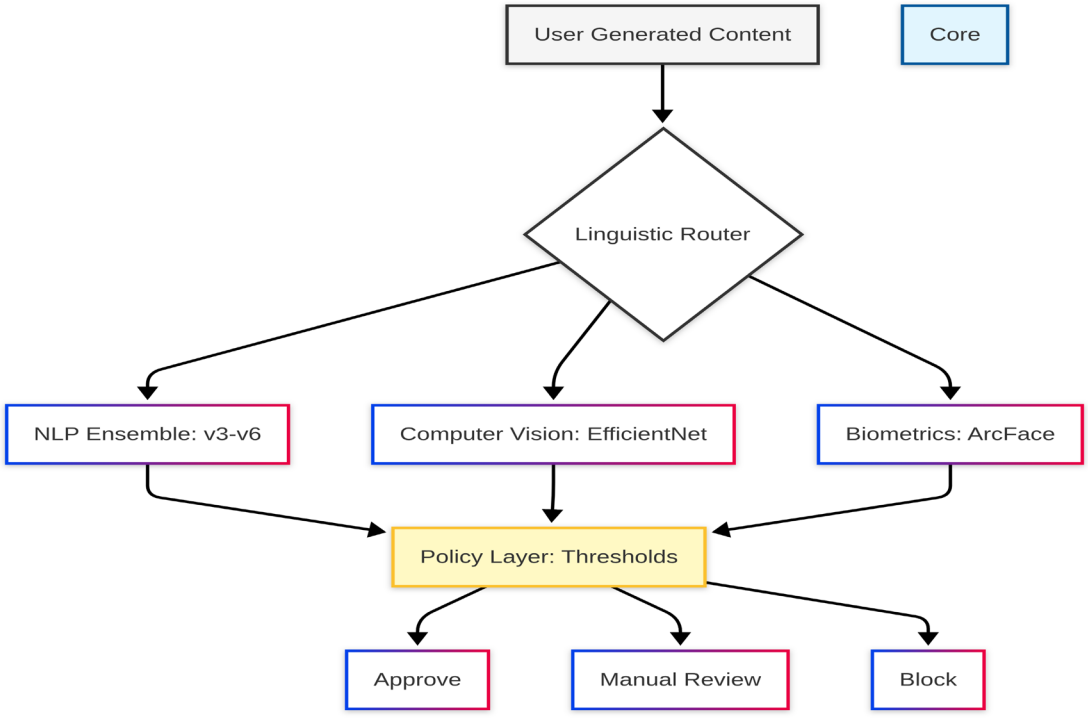

Figure1: The diagram illustrates the asynchronous data flow starting from user-generated content, passing through a linguistic router for model isolation, and reaching the ensemble core (NLP, Computer Vision, and Biometrics). The final decision is governed by the Policy Layer based on calibrated confidence thresholds.

Source: Author's elaboration

As illustrated in Figure 1, the system architecture prioritizes computational efficiency through the Linguistic Router, which ensures that only relevant language models are activated for a given input. The Ensemble Core provides necessary redundancy, where eight models operate in parallel to mitigate single-model classification errors. The final stage, the Policy Layer, acts as the central authority, applying calibrated thresholds to translate raw AI scores into definitive moderation actions: Approve, Manual Review, or Block.

3.5. IDENTIFICATION AND BIOMETRIC MODULE

The system incorporates an embedding-based facial recognition model, ArcFace, to generate compact facial representations for identity verification and vector-based similarity matching. These embeddings enable efficient comparison between facial images while supporting scalable identity-matching operations (Deng et al., 2019).

4.IMPLEMENTATION DETAILS

The system was implemented in Python using PyTorch and Scikit-learn and deployed through Docker containers to ensure reproducibility, environmental isolation, and horizontal scalability. The infrastructure uses Amazon S3 for model-weight storage and dedicated GPU nodes for high-performance asynchronous inference.

4.1. DATASET AND TRAINING PARAMETERS

The visual safety module applies transfer learning using an EfficientNet-B0 backbone pretrained on the ImageNet dataset. To adapt the model to the platform's specific content and operational requirements, it was fine-tuned on a carefully curated custom dataset.

The dataset contains validated samples representing explicit or Not Safe for Work (NSFW) content, violent imagery, and abusive visual patterns, together with benign media collected from the platform. The inclusion of both harmful and non-harmful samples supports more accurate differentiation between policy violations and legitimate user-generated content.

To reduce overfitting and improve generalisation, on-the-fly data augmentation was incorporated into the training pipeline. The applied transformations included random rotations, horizontal flipping, shearing, and brightness adjustments. The model was trained using the Adam optimiser, an initial learning rate of 0.0001, dynamic learning-rate scheduling, and a weighted cross-entropy loss function. Class weights were adjusted to reduce costly classification errors, with particular attention given to limiting false-positive content restrictions.

4.2. NLP ENSEMBLE VOTING LOGIC

To strengthen risk control and reduce false-positive moderation decisions, the policy layer applies a Weighted Confidence Thresholding mechanism rather than a simple majority-voting procedure.

The NLP ensemble consists of eight models operating in parallel. Each model produces an individual confidence score between 0 and 1, where 0 indicates content classified as safe and 1 indicates a high-confidence policy violation. Because the models differ in accuracy and reliability, each is assigned a weight based on its historical precision and validation performance. The sum of all reliability weights is equal to 1.

The aggregated confidence score is calculated as follows:

$$\textit{Aggregated Score} = (\textit{Model 1 Score} \times \textit{Weight 1}) + (\textit{Model 2 Score} \times \textit{Weight 2}) + \ldots + \\ + (\textit{Model 8 Score} \times \textit{Weight 8})$$

where (s_i) denotes the confidence score generated by model (i), and (w_i) represents its assigned reliability weight.

The resulting score determines the appropriate moderation action according to three operational thresholds:

- Approve (Aggregated Score < 0.40): The content is cleared immediately for production deployment as safe.
- Manual Review / Uncertainty Zone (0.40 ≤ Aggregated Score < 0.85): The content is routed to an asynchronous review queue for human verification. This completely mitigates the risk of false censorship and handles borderline cases safely.
- Block (Aggregated Score ≥ 0.85): Immediate autonomous restriction and isolation of the content.

This design ensures that automatic blocking is applied only when the combined outputs of the most reliable models exceed a strict confidence threshold. Consequently, the ensemble reduces dependence on any single classifier, limits isolated model errors, and improves the stability and consistency of the production moderation system.

5. RESULTS AND EVALUATION

The deployment of the custom-tuned moderation architecture significantly outperformed standard baseline configurations. To validate the system's efficiency, the text and visual safety components were benchmarked against their respective standard baselines using standard classification metrics. A primary focus was placed on maximizing Precision to meet the strict operational requirement of minimizing false positives (unjustified user and content blocking).

Table 1. Performance Comparison – Standard Baselines vs. Custom Safety Framework

Metric	Single NLP Baseline	Multi-Model NLP Ensemble	Baseline CNN Classifier	Fine-Tuned CNN Classifier
Accuracy	86.2%	93.8%	84.5%	92.1%
Precision	82.4%	94.5%	81.0%	91.8%
Recall	84.1%	91.2%	83.2%	90.5%
F1-Score	83.2%	92.8%	82.1%	91.1%

Note: Metrics were evaluated on a dedicated, balance-controlled validation benchmark dataset reflecting real-world platform edge cases.

Source: Author's elaboration

As shown in Table 1, model fine-tuning and ensemble-based voting reduced classification errors relative to standard single-model configurations. Precision reached 94.5% for the multimodel NLP ensemble and 91.8% for the fine-tuned CNN classifier, indicating that the Policy Layer's weighted confidence mechanism effectively limited false-positive decisions and improved overall system reliability.

Specialised components, including the ArcFace-based biometric pipeline for facial identification and identity verification, operate as independent filters within the broader architecture. Because these components address biometric matching rather than content-safety classification, their performance is evaluated separately from the moderation metrics reported in Table 1.

Although the parallel execution of multiple models introduces additional computational demands, the distributed asynchronous processing pipeline maintained acceptable end-to-end latency under production-simulated workloads. Moderation decisions were consistently returned within the operational limits required for real-time deployment.

6. CASE STUDY: THE FINDWAY PLATFORM

The proposed system has been deployed within FindWay, an international platform designed to support the search for missing persons, animals, and personal belongings. The platform uses the content moderation engine to enhance user safety and maintain the integrity of submitted information. Its biometric recognition component applies embedding-based facial comparison to identify potential matches between uploaded records. When a sufficiently strong match is detected, the system generates an automated notification for further verification, potentially reducing the time required to identify relevant cases.

7. SOCIO-ECONOMIC IMPACT AND FUTURE WORK

Beyond its technical performance, the system may generate broader social and economic benefits by:

- supporting efforts to locate missing persons;
- facilitating faster responses in urgent and crisis-related situations; and
- improving the safety, operational efficiency, and economic sustainability of digital platforms by reducing the resources required for manual moderation.

Future development will focus on expanding the range of specialised models, improving multilingual and context-sensitive classification, strengthening biometric matching accuracy, and exploring responsible integration with relevant public-safety institutions.

7.1. ETHICAL CONSIDERATIONS AND PRIVACY COMPLIANCE

Given the use of biometric identification, privacy, data protection, and responsible governance are central design considerations. The system is intended to operate in accordance with applicable data-protection requirements, including the principles established by the General Data Protection Regulation where relevant.

The biometric module converts facial features into high-dimensional numerical embeddings used for similarity comparison. Raw facial images are not retained within the biometric matching database after processing, provided that no separate legal or operational requirement justifies their storage. Nevertheless, biometric embeddings remain sensitive personal data because they may enable individuals to be distinguished or identified. Their processing therefore requires an appropriate legal basis, clearly defined retention periods, access controls, encryption, audit mechanisms, and procedures for data-subject rights.

Although embedding-based processing can reduce exposure associated with retaining raw images, it does not eliminate privacy, security, or algorithmic-bias risks. Regular security assessments, demographic performance testing, human verification of potential matches, and transparent governance procedures are therefore necessary to support lawful, fair, and accountable system operation (Hai & Lan, 2026).

8. CONCLUSION

This study demonstrates the practical applicability of a multimodel AI architecture in a high-stakes production environment. By integrating natural language processing, computer vision, ensemble-based decision logic, and biometric embeddings, the proposed system provides a scalable and robust approach to contemporary content moderation and identity-matching challenges. The findings further indicate that distributed processing, model redundancy, confidence-based decision thresholds, and human review can improve system reliability while reducing dependence on individual classifiers. However, the responsible deployment of such systems requires continuous performance monitor-

ing, privacy safeguards, bias assessment, and transparent oversight, particularly when biometric data and automated moderation decisions are involved.

REFERENCES

Bouzaidi, T. D., Daoui, F., Solhi, S., & Tounsi, S. (2026). Economic Convergence in Africa Put to the Test of Productive Complexity: An Empirical Analysis Using System GMM Panel Data and Markov Chains. *ECONOMICS - Innovative and Economic Research Journal, 14*(2), 145-163. https://doi.org/10.2478/eoik-2026-0030

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4690-4699). https://doi.org/10.1109/CVPR.2019.00482

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org/

Hai, T. V., & Lan, T. T. (2026). Determinants of Economic and Corporate Social Responsibility Performance of Securities Companies. *ECONOMICS - Innovative and Economic Research Journal, 14*(2), 165-187. https://doi.org/10.2478/eoik-2026-0031

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Mäkinen, S., et al. (2021). Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? arXiv preprint arXiv:2103.08942. https://doi.org/10.48550/arXiv.2103.08942

Nassanbekova, S., Yeshenkulova, G., Nurguzhina, A., Ibrahim, M., & Ibadildin, N. (2026). Understanding Student Adoption of Generative AI Chatbots in Higher Education: An Integrated TAM–TRI Approach. *ECONOMICS - Innovative and Economic Research Journal, 14*(2), 293-307. https://doi.org/10.2478/eoik-2026-0037

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 6105–6114). http://proceedings.mlr.press/v97/tan19a.html

Vankov, N., & Lukarev, V. (2025). Implementation of smart IoT technologies on the Bulgarian market: An innovative approach for economic sustainability and environmental benefits. *Collection of Papers New Economy, 3*(1), 101–116. https://doi.org/10.61432/CPNE0301101v