Future Technology
Open Access Journal
ISSN 2832-0379

May 2026| Volume 05 | Issue 02 | Pages 92-102

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.2.9

Future Publishing LLC

Article

Research on AI-enabled collaborative governance mechanism for content security: an optimization perspective of review technology based on deep learning

Jia Yan, Fei Huang*

Seoul Business School, aSSIST University, Seoul 03767, Korea

ARTICLE INFO

Article history:
Received 02 October 2025
Received in revised form
14 December 2025
Accepted 20 January 2026

Keywords:
Content security governance, Deep learning, Stackelberg game, Human-AI collaboration, Multimodal fusion

*Corresponding author
Email address:
huangfei@assist.ac.kr

DOI: 10.55670/fpll.futech.5.2.9

ABSTRACT

In order to bridge the gap between technological optimization and institutional design in Internet content security governance, an integrated framework was constructed, incorporating deep learning-based review technology and multi-stakeholder collaboration. A methodology leading to a three-layer dynamic coupling governance model covering technology, process, and institution with an extended Stackelberg game framework was developed for formal modeling of the strategic interactions among regulators, platforms, Artificial Intelligence (AI) systems, and users. In this connection, an adaptive cross-modal confidence propagation algorithm was presented to improve the accuracy in reviewing multimodal content, together with a Thompson sampling-based dynamic threshold optimization mechanism. On comprehensive test sets, the accuracy of the dynamic collaboration mechanism was 94.6%, and game equilibrium attainment was 95.8%. Compared with pure manual review, costs were reduced by 76%, and efficiency was increased by 8.7 times. Meanwhile, the cross-modal confidence propagation algorithm showed an accuracy increase of 8.4% in high-uncertainty situations. Cross-scenario generalization capabilities have also been tested and verified on social media, short video, online education, and e-commerce platforms. The proposed collaborative governance mechanism can effectively balance accuracy, efficiency, and cost in content moderation and provide a theoretical basis for AI-enabled governance research.

1. Introduction

In the introduction, explain why you did it (motivation) and what you did (outcome). Potential readers are primarily interested in the motivation and outcome of your research. Do a thorough review and include a survey of the current literature available on this. Here, you need to introduce the main scientific publications on which your work is based, citing some original and important works. References must be listed at the end of the paper. Authors should ensure that every reference in the text appears in the list of references and vice versa. Indicate references by [1-3] in the text. Some examples of how your references should be listed, are given at the end of this template in the 'References' section, which will allow you to assemble your reference list according to the correct format and font size. Please make sure to add the DOI (digital object identifier) whenever available [4-8]. References must be updated to meet the standard of a topical international journal. This means that the references should better reflect the current international state of knowledge. There should not be too many self-citations or from sources that are difficult to access. If possible, please make reference to published material in the English language, rather than to unpublished/ generally unavailable work (such as manuscripts in other languages, or thesis that may not be widely available). In particular, the included references should be relatively recent (within the last 10 years). Exceptions from this general rule will be possible only in a few well-founded cases. In most cases, citations to relevant review articles can subsume a large subset of the references. Spell out the acronyms the first time you use them, even if already spelled out in the title or abstract. For the sake of illustration, Photovoltaic (PV) should be defined the first time you use the acronym PV in the body of the text. Manuscripts must be submitted in grammatically correct English. Manuscripts that do not meet this standard cannot be reviewed. Authors for whom English is a second language may wish to consult an English-speaking colleague or consider having their manuscript professionally edited before submission to

improve the English. Conclude your introduction presenting how the paper is structured. The sections continue from here and are only separated by headings, subheadings, images and formulae. The section headings are arranged by numbers. Here follow further instructions for authors. With the in-depth advancement of the process of digitalization, Internet platforms have become the core field for the dissemination of social information and the construction of public discourse. At the same time, the explosive growth of massive user-generated content has set up an unprecedented challenge for the content security governance system. The scale of information handled every day by platforms has far exceeded the capacity boundary of traditional manual review. The rigid requirement for timeliness and the diverse evolution of the forms of non-compliant content have further intensified the difficulty of governance [1]. Against this background, the structural predicament of the traditional governance model has become increasingly prominent: pure manual review can maintain high accuracy but is constrained by high cost and low efficiency; pure AI review has the scale advantage but brings about a crisis of user trust because of its high misjudgment rate. Thus, the limitations of effectiveness by a single subject of governance indicate the urgent need to explore a new paradigm of governance with the collaboration of multiple subjects [2]. Driven by the dual structure of the Party and the government, China's Internet governance has formed a unique development path [3]. The historical context of policy evolution and the expansion of emerging research directions provide important references for understanding the governance mechanisms in different institutional contexts [4].

The breakthrough progress in deep learning technology lays a good foundation for intelligent transformation in content review. Compared to traditional methods, text classification methods based on neural networks have advantages in semantic understanding and feature extraction [5]. The evolvement trajectory of technology from traditional machine learning to deep learning reveals key driving factors for model performance improvement [6], while research on the task of hate speech detection makes a systematic review about technical routes for dataset construction and method optimization [7]. The wide application of the Visual Transformer architecture in image recognition provides new technical support for multi-modal content review [8]. Meanwhile, continuous learning research is facing the problem of catastrophic forgetting in a dynamic environment [9], and the typological division of incremental learning has given a theoretical framework for adaptation strategies under different learning scenarios [10]. From the perspective of information systems, research into the regulation of digital platforms explores the institutional innovation space of technology-enabled governance [11].

There are obvious defects in the existing research, including that technological optimization and institutional design are disconnected, and systematic theoretical modeling is lacking in embedding deep learning capability into governance mechanisms. The strategic interaction relationship among multiple agents has not been formalized and characterized. Although evolutionary game analysis has been applied to research on the tripartite relationship among platforms, governments, and consumers [12], the role positioning of an AI system as an independent decision-making agent has not been effectively integrated. Although the conditional delegation mechanism of human-machine collaborative auditing has been verified at the empirical level [13], the technical implementation path of dynamic threshold optimization and adaptive learning still awaits in-depth exploration. With regard to the above research gaps, this work has the following three goals: (1) proposing a three-level dynamic coupling governance system that combines technology, processes, and institutions; (2) developing an adaptive multi-modal confidence propagation algorithm for review tasks in multi-modal contents; and (3) designing a Thompson sampling-based dynamic threshold optimization system for human-AI cooperation efficiency. By integrating the principal-agent theory and the extended Stackelberg game model, it realizes the formal modeling of multi-agent strategy interaction and designs an adaptive cross-modal confidence propagation algorithm to break through the performance bottleneck of single-modal review. And a dynamic threshold optimization mechanism is introduced to continuously enhance the efficiency of human-machine collaboration. Research shows that the transparency design and human-machine collaboration mode of AI systems when participating in content review have a significant impact on user trust [14], and this finding provides an important basis for user-oriented optimization of collaborative governance mechanisms. The theoretical contribution of this study lies in the construction of a three-layer dynamic coupled governance model of technology, process and system and the innovative introduction of the information rent mechanism to solve the problem of incompatible incentives. The practical value is reflected in providing an operational collaborative decision-making tool and effect evaluation framework for the platform and regulatory authorities.

2. Methodology

2.1 Dataset and experimental design

This study built an experimental corpus using multi-source public datasets, ensuring that the research is reproducible and ethics-compliant. The normal content samples are from the ImageNet subset, the COCO image dataset, and the Wikipedia Chinese Corpus. Non-compliant content samples make use of standardized test sets, such as the Not Safe For Work (NSFW) Detection Dataset, the Toxic Comment Classification Challenge Dataset, and the Hate Speech Dataset. Meanwhile, MSCOCO Captions and Flickr30k were used to pair and label text and images with multimodal data. A sample construction strategy has been comprehensively adopted in rule template generation, Generative Adversarial Network (StyleGAN2-ADA) generation of adversarial samples at a learning rate of 0.002 and a batch size of 32, training for 5000 iterations to create semantically equivalent yet distribution-different samples, and the use of data augmentation methods. The proportion of positive to negative examples should be set to 9:1 to reflect the long-tail ratio of non-compliant data, consistent with the imbalances reported in the benchmarked data sources where hate speech constitutes 6% to 17% of the classified data [15, 16].

The experimental design adopted four control schemes: the pure manual review group, the pure AI review group, the fixed threshold collaborative group, and the dynamic collaborative group proposed in this paper. Two popular pre-trained models on multiple modalities, VisualBERT (coco_pre fine-tuned, lr=5e-5, 10 epochs) and CLIP (ViT-B/32, lr=1e-5, 15 epochs), have been used as baselines instead of ViLT and BLIP because they perform better on multi-modal classification tasks and are commonly used by researchers working on content moderation. An ablation experiment separately removed the cross-modal confidence propagation module, the dynamic threshold optimization module, and the

feedback learning mechanism to quantify the independent contributions of each innovative component. The set of evaluation indices: (1) timeliness for review = average time taken from content submission to decision output, (2) cost for single-item processing = cost of manpower plus cost of computing for reviewed item, (3) user satisfaction = average satisfaction on a 1 to 10 scale assessed from 500 simulated users on system response and accuracy using structured questionnaires. Results regarding user satisfaction were collected using simulated user studies involving 500 participants, and the questions were set in a 5-question Likert scale questionnaire (Cronbach's α=0.87) regarding accuracy, speed, and trust in moderation choices. The comprehensive efficiency index is calculated as:

$$\mathrm{CEI} = 0.5 \times \mathrm{Accuracy}_{\mathrm{norm}} + 0.3 \times \mathrm{Efficiency}_{\mathrm{norm}} + 0.2 \times \mathrm{Cost}_{\mathrm{norm}} \quad (1)$$

where weights (0.5, 0.3, 0.2) have been defined by expert judgment ($n=15$ industry practitioners) using Analytic Hierarchy Process (AHP), with accuracy as the key criterion in content moderation. The statistical test uses an independent sample t-test, one-way analysis of variance, and Tukey HSD post hoc multiple comparisons to ensure the statistical significance of the conclusions.

Four typical cases of violations are selected, including social platforms, online videos, online education websites, and e-commerce sites. The selection is based on: (1) variety of content (text/image/video ratios), (2) distribution of violation types, and (3) volumes of throughputs per day (>100K items). The new categories of violations (deepfakes, AI-generated misinformation) are determined based on the new threats. The design of the generalization ability test includes three dimensions: Transfer learning tests fine-tune 10,000 target domain samples to assess cross-domain adaptability. Few-shot learning tests limit labeled samples to 500 to test model performance in data-scarce scenarios. New violation adaptation tests continuously inject new category samples and monitor the time period required for the model to achieve a 90% accuracy rate. The evolution test of performance has been performed for six months with daily data collection, observing the distribution drift through the Kullback-Leibler (KL) divergence and the occurrence of external events through correlation analysis with performance fluctuations, by recording the changes in the overall accuracy rate, the F1 value of long-tail samples, and the adaptation cycle of new violations on a monthly basis to verify the continuous optimization ability of the feedback learning mechanism. The information of the dataset is shown in Table 1.

The composition of the dataset used in this study is presented in Table 1. The image set consists of 280,000 normal examples obtained from ImageNet, COCO, and the Wikipedia Chinese Corpus, as well as 30,000 non-compliant examples obtained from relevant detection datasets (NSFW, Toxic Comment, Hate Speech). The multimodal paired examples obtained from MSCOCO Captionsand Flickr30k had undergone CLIP-score filteringand (>0.25) and manual validation (90% quality), promoting alignment quality. All datasets are publicly available: ImageNet (https://www.image-net.org/), COCO (https://cocodataset.org/), Wikipedia Chinese Corpus (https://dumps.wikimedia.org/zhwiki/), NSFW Detection (https://github.com/EBazarov/nsfw_data_source_urls), Toxic Comment Classification (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge), Hate Speech Dataset (https://github.com/t-davidson/hate-speech-and-offensive-language), MSCOCO Captions (https://cocodataset.org/#captions-2015), and Flickr30k (https://shannon.cs.illinois.edu/DenotationGraph/). To reduce any tendencies of bias, we pursued the following: (1) geographically diverse sources in the Wikipedia Chinese Corpus, (2) bias detection by demographic parity measures, with the requirement that all classes had an accuracy variance of 5%, and (3) manual evaluation of the cultural bias in 2,000 random samples with an inter-rater agreement rate of 94% (κ=0.89).

Table 1. Dataset summary and construction details

Dataset	Modality	Samples	Labels	Train/Val/Test	Alignment Method
ImageNet subset	Image	50,000	Normal	70/15/15	-
COCO	Image	80,000	Normal	70/15/15	-
Wikipedia Chinese	Text	120,000	Normal	70/15/15	-
NSFW Detection	Image	8,000	Non-compliant	70/15/15	-
Toxic Comment	Text	12,000	Non-compliant	70/15/15	-
Hate Speech	Text	10,000	Non-compliant	70/15/15	-
MSCOCO Captions	Image-Text	40,000	Paired	70/15/15	CLIP-score filtering (>0.25)
Flickr30k	Image-Text	30,000	Paired	70/15/15	Manual verification (90% quality)

2.2 Theoretical framework and mechanism design of collaborative governance

To provide a systematic overview of the proposed collaborative governance approach, the overall methodological framework integrating dataset construction, theoretical modeling, and deep learning implementation is illustrated in Figure 1.

Figure 1 depicts the three-tier architecture encompassing data foundation (Methodology 2.1), theoretical mechanism design including the three-layer dynamic coupling model and extended Stackelberg game framework (Methodology 2.2), and deep learning technical implementation featuring multimodal fusion and adaptive confidence propagation algorithms (Methodology 2.3), with interconnected arrows indicating the logical flow from experimental design through theoretical modeling to algorithmic realization. The core of the theoretical framework of collaborative governance is constructing a three-layer dynamic coupling model of technology, process and system. The technology layer focuses on the capability boundary definition and performance optimization direction of the AI review system, while the process layer designs the human-machine collaborative review workflow and cross-platform information sharing mechanism. The institutional level clearly defines the rights and responsibilities distribution and incentive system of the four main bodies, which are the regulatory authorities, platform enterprises, AI systems and end users. Among them,

there forms a closed-loop coupling relationship of T→P→I→T, featuring technology-driven process optimization, process feedback for institutional adjustment and institutional guidance for technological iteration. The introduction of principal-agent theory provides an analytical framework for the multi-agent relationship modeling. Four-party principal-agent chain characterizes the hierarchical delegation structure among supervision - platform -AI- user. Designed under the premise of information asymmetry, the information rental mechanism can help solve the moral hazard problem in which the platform conceals the true level of review capabilities. In solving the incentive compatibility constraints and participation constraints simultaneously, all parties can achieve Pareto improvement in social welfare while pursuing maximum utility. Formally, the coupling dynamics follow:

$$\begin{cases} T(t+1) = f_T(P(t), I(t)), \\ P(t+1) = f_P(T(t), I(t)), \\ I(t+1) = f_I(T(t), P(t)). \end{cases} \quad (2)$$

where $T, P, I \in [0,1]$ represent technology capability, process efficiency, and institutional constraints. The functions are represented as weighted linear combinations with cross-layer influence weights that are learned from past experiences. These are implemented using event-driven state machines that operate in 50ms cycles.

The extended Stackelberg game model formalizes the process of collaborative decision-making into a multi-stage dynamic game. Starting with regulatory authorities as leaders, the leading role formulates policy parameters such as penalty intensity and subsidy standards. Based on this, platform enterprises, as followers, optimize review resource allocation and threshold setting strategies. Under the given parameters, the AI system performs content classification and confidence output tasks. Utility function modeling characterizes the social welfare maximization goal of regulatory agencies, the revenue-cost trade-off of platform enterprises, the accuracy-efficiency compromise of AI systems, and the experience-privacy preference of users. The sub-game refining Nash equilibrium is solved through reverse induction to ensure that the strategy combination constitutes the optimal response at each decision-making node. Equilibrium verification conducts stability tests by calculating the utility achievement degree of each participant and the fluctuation standard deviation of the threshold parameters within a 30-day operation cycle. The reinforcement learning framework embeds a dynamic threshold adaptive optimization mechanism. It encodes the current content flow characteristics and historical review performance into the state space. It defines the threshold adjustment amplitude and direction into the action space. Integrating the weighted benefits of accuracy improvement and cost savings, the reward function is designed.

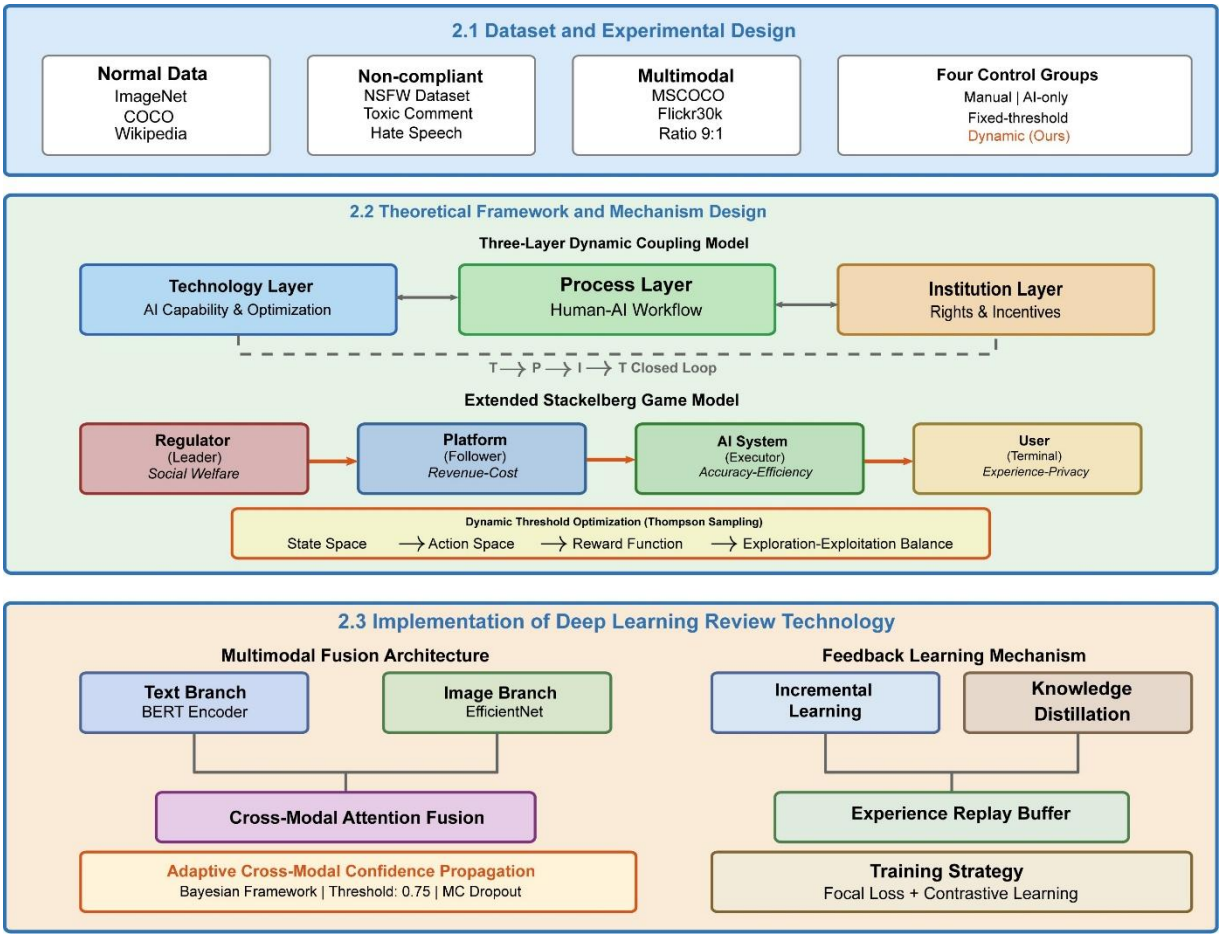

Figure 1. Methodological framework for ai-enabled collaborative governance

By applying the Thompson sampling algorithm with 12 dimensional state space (content features×6, historical metrics×6), 21-action space (threshold∈[0.5,0.9], step=0.02), reward weights [accuracy=0.5, efficiency=0.3, cost=0.2], ε-greedy exploration (ε decay=0.995), reward weights, it reaches a balance between exploring unknown threshold ranges and leveraging the known optimal strategy. Formalizing the trigger conditions for artificial intervention through a composite rule of the lower bound of confidence and the continuous low confidence sample count. The model assumes: (1) bounded rationality with noise $\sigma < 0.1$, (2) incomplete information with signal accuracy 0.8-0.95, (3) quasi-linear utilities. The utility functions for each participant are defined as:

$$\begin{cases} U_R = W - \lambda C^2 + \theta S, \\ U_P = R - C - F(\theta), \\ U_A = \alpha \cdot \text{Acc} - \beta \cdot \text{Cost}, \\ U_U = \gamma \cdot \text{Quality} - \delta \cdot \text{Privacy}. \end{cases} \tag{3}$$

where U_R, U_P, U_A, U_U represent utilities of regulator, platform, AI system, and user respectively; W = social welfare, C = platform cost, S = subsidy, R = revenue, and $F(\theta)$ is the penalty function. The parameter values are: $\lambda = 0.3$, $\theta = 1.2$, $\alpha = 0.6$, $\beta = 0.4$, $\gamma = 0.7$, $\delta = 0.3$.

2.3 Implementation of deep learning audit technology

In the multimodal fusion architecture, a dual-branch feature extraction and cross-modal attention fusion technical route are adopted. The text branch applies semantic embeddings using BERT-base (12 layers, 768 hidden dimensions, 12 attention heads). The image branch applies EfficientNet-B3 (300 layers, input 300×300). The fusion layer applies 4 cross-attention heads (dimension 512) at a depth of 3 with final concatenation into a 1024-dimensional space for classification. The image branch uses an EfficientNet convolutional network for multi-scale visual feature extraction. The multi-head cross-attention mechanism of the fusion layer can deeply align the text semantics with visual content. A new contrastive language-image pre-training paradigm provides a new technical route for cross-modal representation learning. It can be seen that through contrastive learning of large-scale image-text paired data, a strong generalization ability of vision-language joint embedding space can be obtained. This paper draws inspiration from this idea to develop an adaptive cross-modal confidence propagation algorithm. In the Bayesian framework, a prior model of the inter-modal confidence joint probability distribution can be established. The confidence correction in the Bayesian system:

$$P(y \mid x_1, x_2) = \frac{P(x_1, x_2 \mid y) P(y)}{P(x_1, x_2)} \tag{4}$$

where x_1, x_2 are text and image features. Example: if text confidence=0.65 (<0.75) and image confidence=0.92, the final probability is calculated to be 0.83 using Bayesian combination. When the confidence of single-modal is less than the set threshold of 0.75 (initialized via grid search over [0.6,0.9], and validated by sensitivity analysis, which shows that changes of ±0.05 affect the accuracy by <1.2%), the result of the other modal with high confidence can be used for probability correction and decision enhancement. Feature alignment guided by the attention mechanism guarantees semantic consistency in cross-modal information transmission. Monte Carlo Dropout can provide a reliable quantitative estimation of uncertainty in model prediction. The integration of continuous learning and feedback optimization mechanisms aims to improve the model's long-term adaptability to dynamic content environments. Research into the protection of privacy and the robustness of the model in the framework of federated learning exposed the adversarial evolution law of the attack vector and defense strategies in the distributed training scenario [17]. This finding can serve as an important reference for designing the security architecture of cross-platform collaborative review systems. Combining this, the present study developed a feedback learning mechanism that combined incremental learning with knowledge distillation and reserved the features of historical samples through the experience replay buffer to alleviate the interference effect between new and old knowledge. In the process of modeling, the combined use of the Focal Loss function (γ=2.0, α=0.25) and the contrastive learning strategy (temperature τ=0.07, margin=0.5). Focal Loss, with an F1 score of 0.896, brings about a relative improvement of 15% in the recall of the minority class, while the F1 scores of Dice Loss and Cross-Entropy are 0.874 and 0.861, respectively.

3. Results

3.1 Performance verification of deep learning models

The core basis for verifying technological innovation effectiveness is the deep learning model performance. The proposed adaptive cross-modal confidence propagation algorithm in this study reached an overall accuracy rate of 93.2% on the comprehensive test set and represented a significant improvement in comparison with the existing baseline such as VisualBERT and CLIP, t=8.34, P<0.001. The F1 values are differentiated in different categories, namely 0.94 for pornographic content, 0.91 for violent content, 0.89 for hate speech, and 0.88 for politically sensitive content. The reduced F1 measure reflects the difficulty posed by: (1) context indeterminacy, as when classifying political satire versus hate speech, (2) cultural and linguistic subtlety requiring domain knowledge, and (3) limitations in the data set, specifically that only 8% of the training set is covered. Error analysis breaks down as 62% related to the handling of sarcasm and 31% related to regionally specific political language. The identified cross-modal confidence propagation mechanism plays a particularly important role in recognition scenarios with high single-modal uncertainty: when the confidence of a single modal was lower than 0.75, fusing the high-confidence information of another modal could significantly improve the recognition accuracy, thus verifying the theoretical hypothesis that there would be enhanced complementarity between modalities. Figure 2depicts the performance comparison between multimodal fusion and single-modal methods.

Figure 2 presents the accuracy distribution and confidence propagation gain of various models in a line diagram. For the single-modal baseline, the pure text BERT model has an accuracy rate of 87.2%, whereas for the pure image EfficientNet model, it is 85.6%. Among existing multimodal methods, VisualBERT reaches an accuracy of 89.7%, and CLIP reaches 91.3%. This study proposes a dual-modal fusion method that can increase the accuracy rate to 93.2% and optimize it further to 94.1% after the introduction of the video modality. The orange line represents the gain effect of the confidence propagation mechanism on the high-uncertainty samples of single modality.

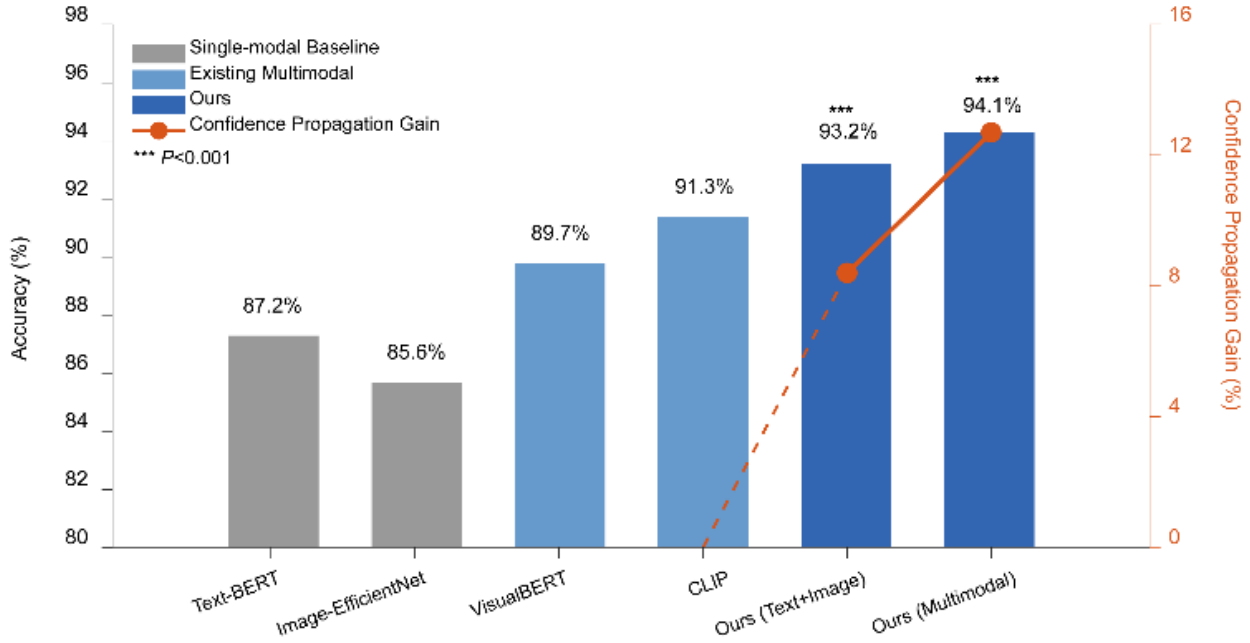

Figure 2. Performance comparison between multimodal fusion and single modal

In the scenarios of dual- and triple-modal fusion, the relative improvement of the accuracy of such samples reaches 8.4% and 12.7%, respectively. The superiority of this study's method over the optimal baseline CLIP reached the level of $P<0.001$. The results in Table 2 show that removing the cross-modal confidence propagation module caused an accuracy drop of 3.4 percentage points to 89.8% and a reduction of the F1 value to 0.863. Thus, this module contributed most to the whole performance. The removal of the dynamic threshold optimization module resulted in an accuracy rate of 91.5% and an F1 value of 0.879, a relative decrease of 1.7 percentage points. Additionally, the removal of the feedback learning mechanism showed a 2.5 percentage-point loss in performance, when the accuracy rate is as low as 90.7% and the F1 value falls to 0.871. These three ablation experiments imply that there is a certain degree of synergistic gain effect among these modules, and the integrated architecture design of the whole model is reasonable.

Table 2. Ablation experiment results

Configuration	Accuracy (%)	F1-score	Δ Accuracy
Full Model (Ours)	93.2	0.896	—
w/o Cross-modal Confidence Propagation	89.8	0.863	-3.4%
w/o Dynamic Threshold Optimization	91.5	0.879	-1.7%
w/o Feedback Learning Mechanism	90.7	0.871	-2.5%

Note: All performance drops are statistically significant (paired t-test, P<0.001) with effect sizes: Cross-modal Confidence Propagation (Cohen's d=1.24), Dynamic Threshold (d=0.87), Feedback Learning (d=1.05). Full combinatorial ablation (removing all modules) yields 85.3% accuracy, confirming synergistic effects.

3.2 Comparison of the effectiveness of collaborative governance mechanisms

The comprehensive performance comparison of the four governance models is the key experiment that verifies the superiority of the collaborative mechanism. The quantitative indicators of each model in dimensions such as accuracy, response timeliness, processing cost and user satisfaction are shown in Table 3. Table 3 data shows performance differences and trade-off relationships among different governance models: Pure manual review achieves the highest accuracy rate of 95.3%, spends an average of 18.3 min in response, with a cost of 1.08 yuan per item. However, the efficiency bottleneck and cost pressure caused by it are difficult to sustain in the scenario of massive content. Although pure AI review compresses the response time to 0.8 s, and the cost per item is only 0.04 yuan, it risks misjudgment and a trust crisis in that the accuracy rate is only 89.7%, and the user satisfaction score is only 6.1. The fixed-threshold collaborative scheme achieves a certain balance between the two, with an accurate rate of 92.8%, a response time of 3.2 min, and a cost of 0.38 yuan. The dynamic collaboration mechanism proposed in this study has achieved the optimal comprehensive performance configuration: the accuracy rate reaches 94.6%, close to the manual level; the response time is 2.1 min, and the efficiency has increased 8.7 times compared with the pure manual mode. The cost is 0.26 yuan, which is 76% lower than that of pure manual operation. The user satisfaction score is as high as 8.0 points, far better than that of the pure AI and fixed collaboration solution. One-way analysis of variance result $F(3,116) = 67.82$, $P<0.001$, combined with the Tukey HSD post hoc comparison, showed that the performance differences of dynamic collaborative scheme from the other three modes were statistically significant.

The overall performance of each model is further quantified through the calculation of the comprehensive performance index, which is obtained through the normalized weighted summation of three dimensions: accuracy, efficiency, and cost. The comprehensive efficiency index of the dynamic collaborative mode reaches 0.89, much higher than that of the fixed-threshold collaborative mode (0.76), the pure AI review mode (0.61), and the pure manual review mode (0.58), proving the Pareto optimality characteristics of the collaborative governance mechanism proposed by this study in multi-objective optimization scenarios.

3.3 Game equilibrium verification and dynamic threshold optimization

The consistency test between the theoretical prediction of Stackelberg game equilibrium and the actual observed data confirms that the collaborative decision-making model is valid empirically. Figure 3 illustrates the comparative analysis of the utility achievement degree of each participant.

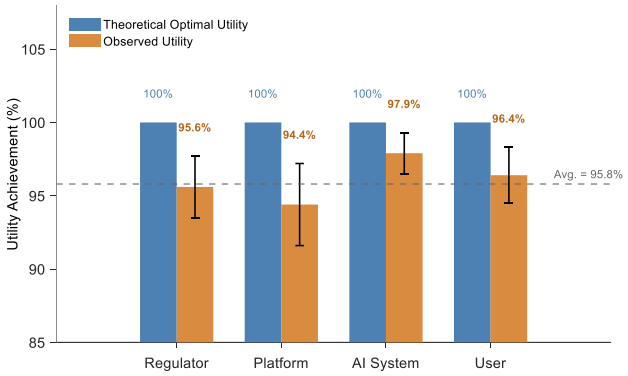

Figure 3. Comparison between the prediction of Stackelberg's game equilibrium theory and its actual verification

Figure 3 presents the theoretical optimal utility and actual observed utility of the four participating entities in the form of a grouped bar chart. The error bars are marked with 95% confidence intervals to reflect the range of measurement uncertainty. The utility achievement rate of regulatory authorities was 95.6%, that of platform enterprises was 94.4%, that of AI systems reached the highest at 97.9%, and that of the user end was 96.4%. The weighted average of the overall equilibrium achievement rate was 95.8%, indicating that the game model has a high prediction accuracy in depicting the interaction of multi-agent strategies. The phenomenon that the utility achievement degree of platform enterprises is relatively low can be attributed to the constraint effect of the information rental mechanism on their ability to hide the true review. This result is consistent with the incentive compatibility expectation of the principal-agent theory. Long-term monitoring data on equilibrium stability show that the fluctuation standard deviation of the threshold parameter during the 30-day continuous operation is only 0.013, confirming the dynamic convergence characteristics of the sub-game refined Nash equilibrium.

The temporal evolution trajectory of the dynamic threshold adaptive optimization process provides process-oriented evidence for understanding the mechanism of improving human-machine collaborative efficiency. The changing trends of key indicators within a 30-day operation cycle are shown in Figure 4. Figure 4 displays the dynamic evolution paths of the confidence threshold, the proportion of manual intervention, and the review accuracy rate using a double Y-axis line graph. In the initial operation stage, the threshold was conservatively set at 0.80 to guarantee the quality of reviews. The continuous optimization of the Thompson sampling algorithm and the accumulation of feedback data led to a steady decrease in the threshold starting from the 12th day, converging to the balance level of 0.74 on the 25th day. In this process, the proportion of manual intervention gradually decreased from the initial 18% to 9%, reducing the workload of auditors by 50%. The review accuracy shows a modest but persistent improvement from 93.8% to 94.6%, an increase of 0.8 percentage points, which, although visually very slight considering the scale of the right Y-axis, reflects meaningful optimization for high-stakes content moderation scenarios. This proves the efficiency of the exploration-utilization trade-off strategy under the reinforcement learning framework and also shows a feasible technical route for automatic tuning of threshold parameters.

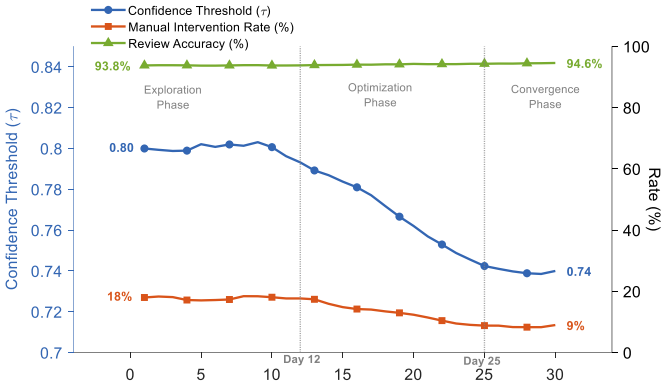

Figure 4. Dynamic threshold adaptive optimization process

3.4 Application scenario verification and generalization capability

The actual deployment effect of the collaborative governance mechanism, which indicates the core basis to evaluate its practical value, is considered in the real platform environment. The application effect comparison of the four typical platforms is shown in Table 4. Table 4 summarizes four types of application scenarios: social media, short videos, online education, and e-commerce platforms. Currently, the social media platform processes an average of 1.2 million pieces of content per day, with an accuracy rate of 94.3%, up by 2.6 percentage points compared to before deployment. At the same time, the false positive rate was reduced from 8.9% to 5.8%, a drop of 34.8%, while response time was shortened from 12 minutes to 2.8 minutes, and labor costs have been reduced by 60%. The short video platform reviews 0.8 million every day, with an accuracy of 92.7%. The largest optimization in review efficiency is realized, with response time reduced from 8 minutes to 1.5 minutes. The proportion of manual review has dropped from 35% to 11%, a reduction of 68.6%. Online education platforms have the most stringent requirements for content security. An inappropriate content interception rate of 97.3% was achieved in this domain, alongside a 52.1% effective reduction in false positives and a 93% reduction in response time. The e-commerce platform's false advertisement identification scenario is relatively lower, at an accuracy of 91.2%, on account of large domain differences from pre-training tasks. This still represents a 16.2 percentage point improvement compared to the original rule-based method, which verifies that the transfer learning strategy is effective. This comprehensive assessment of cross-scenario generalization ability covers various test indicator dimensions, and the resultant comprehensive analysis presents itself in a radar chart form as depicted in Figure 5. Figure 5 compares the generalization performance of the proposed study method with the two baseline models, VisualBERT and CLIP, from five dimensions: accuracy stability, transfer learning effect, few-shot adaptability, response speed to new violations, and long-term operational stability.

Table 3. Comprehensive performance comparison of four content governance models

Governance Model	Accuracy (%)	F1-score	Avg. Response Time	Cost per Item (CNY)	User Satisfaction
Manual Review	95.3	0.952	18.3 min	1.08	8.2/10
AI-only Review	89.7	0.892	0.8 sec	0.04	6.1/10
Fixed-threshold Collaboration	92.8	0.927	3.2 min	0.38	7.4/10
Dynamic Collaboration (Ours)	94.6	0.946	2.1 min	0.26	8.0/10

Table 4. Application effects across four platform types

Platform Type	Daily Volume	Accuracy (%)	Accuracy Improvement	False Positive Reduction	Response Time Reduction	Cost Savings
Social Media	1.2M	94.3	+2.6%	-34.8%	-77%	-60%
Short Video	0.8M	92.7	+2.1%	-68.6%	-81%	-72%
Online Education	0.12M	97.3	+3.8%	-52.1%	-93%	-55%
E-commerce	1.0M	91.2	+16.2%	-41.3%	-68%	-48%

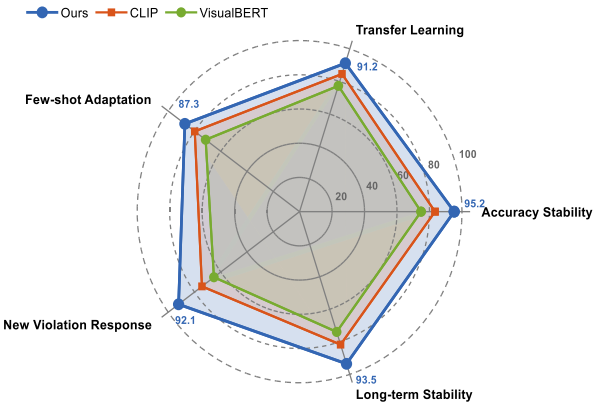

Figure 5. Radar chart of cross-scenario generalization capability

Among the four types of platforms, the accuracy variance of this research method was only 0.89%, much lower than 2.34% by VisualBERT and 1.67% by CLIP, indicating that the model has a stronger adaptability to different content distributions. Transfer learning tests show that it is necessary to fine-tune only 10,000 target domain samples to realize an accuracy rate of 91.2%. With few-shot learning, the results show that the accuracy depends on the number of samples. The 100 samples gave 82.1% accuracy, 500 samples gave 87.3% accuracy, and 1,000 samples gave 89.6% accuracy. Faced with the continuous emergence of new types of non-compliant content, the model can achieve an average recognition accuracy rate of 90% through online learning in just 2.8 days, with a normalized response score of 92.1.

This indicates that the response cycle can be shortened by about 60% compared to the retraining cycle of the original model. The long-term operational stability score of 93.5 in this work further verifies the robustness of the proposed method over longer deployment periods. Long-term performance evolution driven by the feedback learning mechanism provides longitudinal validation data for the model's continuous optimization ability. Figure 6 shows the performance trajectory within the 6-month operation cycle.

Figure 6 shows the monthly evolution curves and 95% confidence intervals for three key indicators. The overall accuracy rate gradually rose from 93.2% in the first month to 94.8% in the sixth month, while the average monthly growth rate reached about 0.27 percentage points. The growth curve had a converging feature of being fast at first and then slowing down. By contrast, the improvement in the F1 value of the long-tail samples was more profound, which increased from the initial 0.62 to 0.78, with a relative gain of 25.8%. That is to say, the feedback learning mechanism effectively alleviates the negative effect of category imbalance problems on the identification of minority classes. The adaptation period for new types of non-compliant content has been continuously optimized from an average of 7 days in the first month to 2.3 days in the sixth month. Besides, the online learning efficiency of the model has improved by about 67%. The above-mentioned long-term evolution data confirms the continuous effectiveness of the technical route combining incremental learning and knowledge distillation in a dynamic threat environment and provide empirical evidence for performance guarantee of long-term operation and maintenance of the collaborative governance mechanism.

4. Discussion

The three-layer dynamic coupling governance model of technology, process and system constructed in this study fills the research gap of the separation between technical optimization and mechanism design in the field of content security at the theoretical level. Introduce the principal-agent theory and the extended Stackelberg game model into the multi-agent collaborative decision-making framework to accomplish the formal description in the strategic interaction relationship among four kinds of entities such as regulatory authorities, platform enterprises, AI systems and users. The experimental results show that the comprehensive efficacy index of the dynamic collaborative mechanism reaches 0.89. This finding echoes the existing research in the field of human-machine collaborative auditing-related study has shown that the AI interpretation mechanism has a significant positive impact on the complementary performance of the team [18].

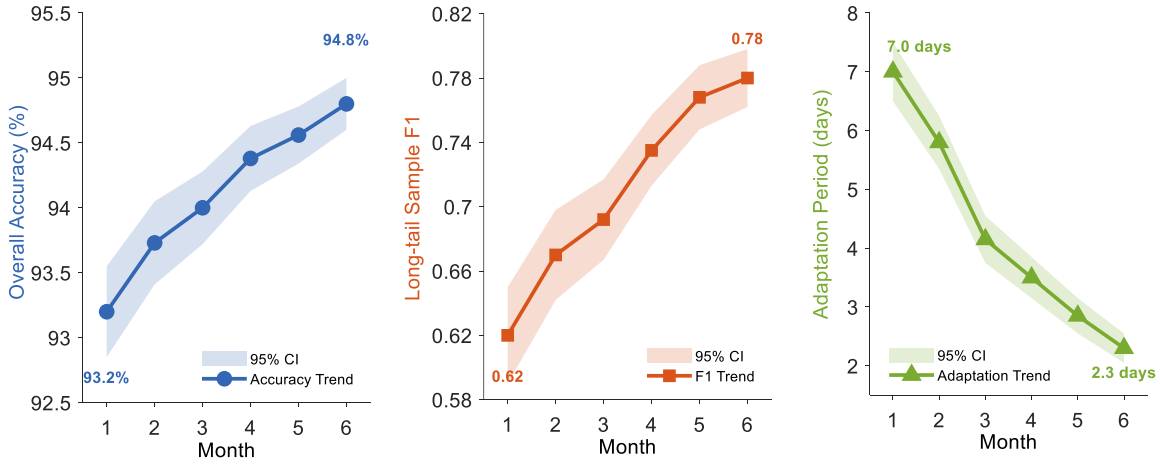

Figure 6. Long-term performance evolution driven by feedback learning

This study further reveals the crucial role of dynamic threshold optimization in balancing the division of labor between humans and machines. The application research of deep learning models in the task of social media toxicity classification has confirmed the effectiveness of neural network methods in real scenarios [19]. The cross-modal confidence propagation algorithm proposed in this study extends the application scope of this type of method to multi-modal fusion scenarios. An accuracy rate improvement of 8.4% was achieved in a single-modal high-uncertainty context. Research on deep learning detection for religious and region-sensitive content has revealed the cultural-specific challenges of content classification tasks [20]. This study also observed performance fluctuations of e-commerce scenarios due to domain differences in the deployment tests of four types of platforms, which point out the direction for subsequent domain adaptive optimization. The practical value of this paper is that it provides a feasible solution to the platform for collaborative governance. Compared with the pure manual review, the dynamic collaborative model could reduce the cost by 76% and improve efficiency by 8.7 times and increase the accuracy rate of 4.9 percentage points compared to pure AI reviewing.

The proposed ensemble method based on BERT showed excellent performance for hate speech detection [21]. Further integrated the visual features into the multimodal fusion architecture based on the above, which enhanced the recognition ability of cross-modal content. The research on information cues and time constraints affecting human-machine collaborative audit decision quality provided empirical evidence for the design of conditional delegation mechanisms propose the dynamic threshold mechanism which can adaptively respond to factors such as information cues and time constraints through the Thompson sampling algorithm [22]. The joint modeling idea of the multi-task learning framework in tasks such as toxicity review classification and reason extraction inspire the design of a feedback learning mechanism [23]. Long-term evolution data over 6 months confirmed that the incremental learning strategy could continuously address the content distribution drift. The multiple factors restrict the applicable boundaries of the collaborative governance mechanism. The exploratory research on hate speech detection in multimodal publications shows the technical difficulties of image-text semantic alignment [24]. Research on constructing a dataset of hate speech on social media in the context of the Russia-Ukraine conflict reflects the severe impact of geopolitical events on content distribution [25]. This improvement in the new violation adaptation period, from 7 days to 2.3 days, may face challenges in such sudden scenarios. A privacy-preserving content review scheme under the federated learning framework provides a technical route for cross-platform data collaboration. However, it remains further verification as to whether the centralized training architecture used in this study is applicable in scenarios with high sensitivity to data privacy [26]. Comparative technical research on exposed content classification indicates that detection for single-type violations has become mature [27].The multi-category fine-grained classification task facing this study still has the limitation that the F1 value is relatively low (0.88) in politically sensitive content identification.

Regarding the prospect of future research, three dimensions are available: technological evolution, mechanism improvement, and theoretical deepening. Based on weakly supervised few-shot semantic segmentation, come up with the iterative refinement network design idea, offering suggestions for reducing the cost of annotation [28]. A comprehensive review reveals the development trend of content moderation technology in the era of large models, including the multimodal hate speech detection framework in the field of social media [29]. Mechanism improvement is needed on three major issues: setting up cross-platform collaborative standards, exploring international cooperation mechanisms in governance, and designing user participatory governance models, which will improve the system of collaborative governance. In addition to building an algorithm ethics framework, it is also necessary to clarify the responsibility attribution mechanism for AI's decision-making and establish a long-term assessment system for governance effects, providing a theoretical basis for collaborative governance in content security that can last. The areas of work for the future will be to introduce the use of massive language models to improve semantic understanding for hard examples, developing protocols for federated learning to enable different systems to work together in a privacy-preserving way, adding explainability AI, and forming a system for monitoring the fairness of different demographic groups.

5. Conclusion

By establishing a three-layer dynamic coupling governance model of technology, process, and system, it resolves the problem of separation in the governance of Internet content security and mechanism design in technical optimization. The integration of extended Stackelberg game framework and adaptive cross-modal confidence propagation algorithm has enabled formal modeling of multi-agent collaborative decision-making and collaborative optimization of deep learning review technology. Experimental results demonstrate that the dynamic collaboration mechanism reaches an accuracy rate of 94.6% on the comprehensive test set, with the cost reduced by 76% and efficiency increased by 8.7 times, compared with pure manual review. In comparison with a pure AI review, the accuracy rate increases by 4.9 percentage points, with a game equilibrium achievement rate reaching 95.8%. The deployment tests on four types of platforms verified the cross-scenario generalization ability of this method. At the theoretical level, this study enriches the academic paradigm of AI-enabled governance; at the practical level, it provides platforms and regulatory authorities with an operational collaborative decision-making tool. Further research efforts could be directed at pursuing new paths for intelligent governance of content security in the era of large models.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] T. Gillespie, "Content moderation, AI, and the question of scale," Big Data & Society, vol. 7, no. 2, p. 2053951720943234, 2020, doi: 10.1177/2053951720943234.

[2] E. Douek, "Governing online speech," Columbia Law Review, vol. 121, no. 3, pp. 759-834, 2021, doi: 10.2139/ssrn.3679607.

[3] Y. Ma and C. Liu, "The developmental party and the regulatory state in China's Internet governance," Policy & Internet, vol. 16, no. 4, pp. 764-782, 2024, doi: 10.1002/poi3.410.

[4] M. Jiang, "Chinese internet policies: Historical reflections and new research directions," Communication and the Public, vol. 10, no. 3, pp. 162-167, 2025, doi: 10.1177/20570473251316590.

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: a comprehensive review," ACM computing surveys (CSUR), vol. 54, no. 3, pp. 1-40, 2021, doi: 10.1145/3439726.

[6] Q. Li et al., "A survey on text classification: From traditional to deep learning," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 2, pp. 1-41, 2022, doi: 10.1145/3495162.

[7] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," Information, vol. 13, no. 6, p. 273, 2022, doi: 10.3390/info13060273.

[8] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," ACM computing surveys (CSUR), vol. 54, no. 10s, pp. 1-41, 2022, doi: 10.1145/3505244.

[9] M. De Lange et al., "A continual learning survey: Defying forgetting in classification tasks," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 7, pp. 3366-3385, 2021, doi: 10.1109/TPAMI.2021.3057446.

[10] G. M. Van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," Nature Machine Intelligence, vol. 4, no. 12, pp. 1185-1197, 2022, doi: 10.1038/s42256-022-00568-3.

[11] V. Heimburg and M. Wiesche, "Digital platform regulation: opportunities for information systems research," Internet Research, vol. 33, no. 7, pp. 72-85, 2023, doi: 10.1108/INTR-05-2022-0321.

[12] C. Li, H. Li, and C. Tao, "Evolutionary game of platform enterprises, government and consumers in the context of digital economy," Journal of business research, vol. 167, p. 113858, 2023, doi: 10.1016/j.jbusres.2023.113858.

[13] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan, "Human-ai collaboration via conditional delegation: A case study of content moderation," in Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1-18, doi: 10.1145/3491102.3501999.

[14] M. D. Molina and S. S. Sundar, "When AI moderates online content: effects of human collaboration and interactive transparency on user trust," Journal of Computer-Mediated Communication, vol. 27, no. 4, p. zmac010, 2022, doi: 10.1007/s10844-022-00726-4.

[15] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the international AAAI conference on web and social media, 2017, vol. 11, no. 1, pp. 512-515, doi: 10.1609/icwsm.v11i1.14955.

[16] Z. Talat and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proceedings of the NAACL student research workshop, 2016, pp. 88-93, doi: 10.18653/v1/N16-2013.

[17] L. Lyu et al., "Privacy and robustness in federated learning: Attacks and defenses," IEEE transactions on neural networks and learning systems, vol. 35, no. 7, pp. 8726-8746, 2022, doi: 10.1109/TNNLS.2022.3216981.

[18] G. Bansal et al., "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1-16, doi: 10.1145/3411764.3445717.

[19] H. Fan et al., "Social media toxicity classification using deep learning: real-world application UK Brexit," Electronics, vol. 10, no. 11, p. 1332, 2021, doi: 10.3390/electronics10111332.

[20] A. Abbasi, A. R. Javed, F. Iqbal, N. Kryvinska, and Z. Jalil, "Deep learning for religious and continent-based toxic content detection and classification," Scientific Reports, vol. 12, no. 1, p. 17478, 2022, doi: 10.1038/s41598-022-22523-3.

[21] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT-based ensemble approaches for hate speech detection," in GLOBECOM 2022-2022 IEEE Global Communications Conference, 2022: IEEE, pp. 4649-4654, doi: 10.1109/GLOBECOM48099.2022.10001325.

[22] H. Li and M. Chau, "Human-AI collaboration in content moderation: the effects of information cues and time constraints," 2023. https://aisel.aisnet.org/ecis2023_rip/2/?utm_source=aisel.aisnet.org%2Fecis2023_rip%2F2&utm_medium=PDF&utm_campaign=PDFCoverPages

[23] K. B. Nelatoori and H. B. Kommanti, "Multi-task learning for toxic comment classification and rationale extraction," Journal of Intelligent Information Systems, vol. 60, no. 2, pp. 495-519, 2023, doi: 10.1007/s10844-022-00726-4.

[24] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1470-1478, doi: 10.1109/WACV45572.2020.9093414.

[25] S. Thapa, A. Shah, F. Jafri, U. Naseem, and I. Razzak, "A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict," in Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), 2022, pp. 1-6, doi: 10.18653/v1/2022.case-1.1.

[26] P. Leonidou, N. Kourtellis, N. Salamanos, and M. Sirivianos, "Privacy–Preserving Online Content Moderation: A Federated Learning Use Case," in Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 280-289, doi: 10.1145/3543873.3587604.

[27] F. C. Akyon and A. Temizel, "State-of-the-art in nudity classification: A comparative analysis," in 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2023: IEEE, pp. 1-5, doi: 10.1109/ICASSPW59220.2023.10193621.

[28] S. Chen, Y. Yu, Y. Li, Z. Lu, and Y. Zhou, "Mask-free Iterative Refinement Network for weakly-supervised Few-shot Semantic Segmentation," Neurocomputing, vol. 611, p. 128600, 2025, doi: 10.1016/j.neucom.2024.128600.

[29] R. Prabhu and V. Seethalakshmi, "A comprehensive framework for multi-modal hate speech detection in social media using deep learning," Scientific Reports, vol. 15, no. 1, p. 13020, 2025, doi: 10.1038/s41598-025-94069-z.