International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

AI-Powered Influencer Video Auto-Posting Platform with Smart Captioning, Hashtag Generation, and Fake Video Detection: A Review

Aryan Raj, Ayush Srivastava, Aparna Sivaraj
Sharda School of Engineering and Technology, Sharda University, Knowledge Park III, Greater Noida, 201306, India
Aryanrajpanki@gmail.com, Srivastavaayush715@gmail.com, aparna.sivaraj@sharda.ac.in

Abstract

The rapid growth of user-generated multimedia content on the web has amplified the need for automated systems to generate captions, personalise hashtags, and detect deepfakes. Not only do these technologies increase user interaction, but they also play a significant role in protecting trust and authenticity online. Previous surveys have widely studied these areas separately. Captioning studies of image reviews have progressed from CNN-RNN models to attention and transformer models, and have mostly optimised for fluency and coherence. Hashtag recommendation research focuses on personalisation approaches, often using user behaviour modelling to enhance content exposure. Likewise, deepfake detection surveys point to multimodal fusion and adversarial learning methods to enhance detection strengths. Though these contributions are noteworthy, they tend to be domain-specific, failing to account for cross-task dependencies and real-world deployment issues. Additionally, repeated constraints, such as English-language bias in datasets, poor handling of low-resource settings, and inadequate attention to trustworthiness and ethics, limit their generalizability. This review addresses these limitations by bringing research in captioning, hashtag recommendation, and deepfake detection under a single umbrella of social media automation. By systematically connecting generation, personalisation, and verification, it delineates common challenges such as dataset bias, real-world generalisation, and scalability, while charting bleeding-edge solutions that aim to span these gaps. In contrast to previous disjointed surveys, this review focuses on the interaction between technical design and platform credibility, offering a unifying perspective. Finally, it lays the groundwork for future research that progresses not just algorithmic performance but also the ethical, reliable incorporation of AI-based multimedia systems into digital ecosystems.

Keywords: *Image captioning, Hashtag recommendation, Deepfake detection, Attention mechanism, Transformer models, Social media automation*

I. Introduction

Artificial intelligence (AI) has come to the forefront of content creation, personalization, and verification on social media platforms. From image captioning and hashtag suggestions to deepfake detection, not only do they improve the user experience, but they also help protect trust in digital ecosystems. Among them, image captioning connects computer vision and natural language generation, allowing for automatic generation of semantically aligned text descriptions of images Matteo et al. [2][3]. Even with fast progress, work in these fields has largely evolved in silos. Image captioning concerns semantic consistency across modalities, hashtag suggestion concerns discoverability, and deepfake detection concerns authenticity and trust. Practically speaking, such dimensions intersect within social media as generation, personalization, and verification processes together mold user interaction. A disconnected perspective may miss out on synergies and platform-level effects.

In image captioning, Asmaa et al. found that attention mechanisms improved BLEU and METEOR scores relative to non-attention systems, demonstrating the effectiveness of context-sensitive feature alignment. Likewise, hashtag recommendation techniques have evolved from frequency-based systems to deep learning. Chen et al. introduced a personalised multimodal model, with impressive gains in precision and recall, while trending-topic approaches [4] stress flexibility with respect to real-time topics. Detection of deepfakes has also improved, with multimodal questionnaires [10][12][18] indicating improvement in the combination of audio-visual cues, although cross-dataset generalisation is still a significant challenge. While surveys of captioning from [1] to [3], hashtag suggestion [4] to [6], and deepfake detection [10] to [18] are extensive within their individual areas, they seldom combine views across tasks. Analogously, policy-related works [19][20] consider governance but not technical harmony under automated pipelines. Inadequate benchmarking across diverse settings and limited investigations into trust-building mechanisms also limit applicability. This review fills these gaps by bringing captioning, recommendation, and detection research together under a shared governance and automation framework. In contrast to previous stand-alone surveys, it highlights the interrelationship between algorithmic innovation and societal trust and offers a unified roadmap for next-generation AI-based social media spaces.

II. Literature Review

Review on those who don't have quantitative results:

From Show to Tell: A Survey on Image Captioning (2018) [1] presents one of the first integrated reviews of image captioning models, summarizing the transition from CNN-RNN pipelines to encoder-decoder frameworks. CNNs were predominantly applied to feature extraction while RNNs produced sequential captions. MS COCO and Flickr30k were mentioned as dominant benchmarks. Its value comes in classifying template based, retrieval based, and generative approaches and demonstrating how deep learning made more coherent and human-like captions possible. No new accuracy results were reported, but the review highlighted the advantages of deep models over previous hand crafted ones. A Survey on Attention-Based Models for Image Captioning (2019) [2] focused only on attention mechanisms and how these provided better caption quality compared to standard CNN-RNN models. Variants like adaptive and spatial attention were examined for their capacity to bridge words with visual areas. MS COCO and Flickr30k datasets were once again the focus, while BLEU, METEOR, and CIDEr were mentioned as standards for evaluation. The primary contribution was demonstrating how attention always improved contextual alignment and caption fluency, although the paper did not present novel experimental results.

Unmasking Deepfakes: A Review of Datasets, Tools & Techniques (2021) [13] reviewed deepfake detection that includes CNN based classifiers, recurrent networks for temporal features, and multimodal fusion. It listed major datasets, including FaceForensics++, Celeb-DF, and DFDC. The contribution was a methods-and-tools taxonomy, and the results showed that, despite numerous models achieving good within-dataset performance, cross-dataset generalisation is poor. Lastly, Human Performance in Detecting Deepfakes: Systematic Review (2022) [19] integrated research on how humans judge deepfakes. Findings indicated that individuals often mislabel manipulated information, even with warnings, doing much worse than computer programs. Its major contribution is the need to integrate machine detection with human-centred design for authentic content moderation, as more fully described in Table 1.

Table 1: Literature review

Reference	Year	Algorithm / Model	Dataset	Contributions	Results
[17]	2024	Benchmark dataset	Social media content	New large-scale benchmark	Includes baseline AUC=0.82, F1=0.74.
[9]	2023	Transformer-based generation	Microblogs	Generates hashtags based on text segments.	BLEU-4≈28.4, ROUGEL≈ 33.1
[15]	2023	Fusion strategies	Multiple benchmarks	Single vs multimodal fusion	Fusion strategies boosted AUC to ~0.97.
[8]	2022	Graph embeddings	Low-resource social media	Hashtag personalization in low-resource scenarios	F1-score improved by 9% in multilingual
[3]	2022	Deep learning (CNN, RNN, Transformer)	Flickr30k, COCO, custom	Consolidated review of datasets and evaluation	Summarizes SoTa BLEU≈40, CIDEr≈ 120 in transformer models.
[12]	2022	Hybrid DL methods	FaceForensics++, CelebDF	Detailed survey of detection approaches.	Summarizes performances: AUC0.89-0.97 across models
[14]	2022	Multimodal autonomous approaches	Benchmark datasets	Broad multimodal scan	Reports accuracy range 80–95% depending on dataset.
[6]	2021	Multimodal personalization	Twitter + demographic info	Personalized hashtag suggestions.	Improved Recall @10 over baselines by ~6%.
[10]	2021	CNN, RNN, Transformer fusion	FaceForensics++, CelebDF	Focus on cross-dataset robustness	Accuracy drops~20% when trained/tested cross-dataset.
[16]	2021	Deep learning comparison	Face datasets	Comparison of deepfake detection methods.	Best model accuracy ≈ 91.5% on Celeb-DF.
[18]	2021	Human study	Mixed datasets	Analyzes human detection performance.	Human accuracy ≈ 50–60%.
[11]	2020	Multimodal DL approaches	Multiple public datasets	Survey of detection and generation	Covers AUC,EER benchmarks; notes SoTA AUC≈ 0.95
[7]	2020	Hybrid visual model	Instagram images	Fusion improves accuracy	Reported MAP@10=0.41.

[4]	2020	Dynamic neural model	News dataset	Hashtag personalization adapting to trending topics.	Reported Precision@5≈0.68, Recall@5≈0.52
[5]	2019	End-to-end deep learning	Twitter dataset	Automatic hashtag generation from tweets.	Accuracy≈72%, F1-score ≈0.63.
[20]	2019	AI-driven scheduling system	Simulated social media data	Discusses architecture for automated scheduling.	Reports engagement improved by ~18%.

A. Image Captioning Surveys

Recent image captioning surveys like Hossain et al. (2019) [1] and Asmaa et al. (2019) [2] are primarily focused on algorithmic developments, ranging from initial CNN-RNN pipelines to increasingly sophisticated attention-based and transformer models. Hossain et al. highlighted the general taxonomy of models, whereas Asmaa et al. indicated how attention mechanisms enhance BLEU and METEOR scores over non-attention models. While these reviews are informative regarding the evolution of captioning models, they do not put as much emphasis on cross domain applications, e.g. integration with recommendation systems or misinformation detection pipelines. Another limitation is dataset bias: the majority of discussed datasets are English and high resource biased, with very little focus on multilingual or low-resource cases. Moreover, evaluation procedures center on semantic relevance and fluency at the expense of ethics and trustworthiness, which are paramount for social media usage.

B. Auto Hashtag generation Surveys

For hashtag suggestion, Sharma et al. (2020) [4] highlighted personalisation, whereas Rizwan et al. (2021) [5] provided end-to-end deep learning approaches specific to Twitter. In Table 1 section (B) These works highlight the importance of tailoring recommendations to user needs and habits. Nonetheless, they also identify significant lacunae. The majority of the research has not examined the magnitude of the role of hashtags in propagating misinformation or malicious content. Moreover, there is no cross-platform benchmarking: techniques are often customised to Twitter, and Instagram, TikTok, or multilingual networks are underdeveloped. Another deficiency lies in the assumptions about stable user behaviour. Personalisation models generally hold up under regular patterns but fail under rapid cultural and social changes.

C. Deepfake Detection Surveys

Deepfake detection studies like Nguyen et al. (2019) [10] and Tolosana et al. (2020) [12] investigate multimodal fusion strategies that integrate video, audio, and text signals to enhance accuracy. Though these studies show robust algorithmic development, they keep emphasising the problem of weak cross-dataset generalisation. Models learned on a single dataset tend to do poorly when transferred to another, prompting the issue of real-world relevance. The reviews recognize advances in feature engineering, neural architectures, and multimodal analysis, but also point out outstanding issues of robustness and adaptability in the face of new manipulation methods.

Through Table 1, recent image captioning, hashtag suggestion, and deepfake detection surveys offer comprehensive but isolated perspectives. Captioning reviews inform model development but do not consider ecosystem level implications (Hossain et al. [1], Asmaa et al. [2]). Hashtag surveys emphasise personalisation while downplaying trust and platform-level robustness (Sharma et al. [4], Rizwan et al. [5]). Deepfake detection review papers focus on multimodal solutions but are neither generalizable nor deployable (Nguyen et al. [10], Tolosana et al. [12]). Finally, policy-focused papers discuss compliance while excluding integration into automation pipelines [19], [20]. Collectively, these gaps underscore the urgent need for a comprehensive review that situates technical progress within a social media automaton and trustworthiness framework. By interrelating generation, personalization, and verification, your review not only charts cutting-edge approaches but also establishes foundations for future research spanning technical design, platform governance, and societal trust.

III. Methodology

The methodology of the review follows a structured approach to identifying, classifying, and analysing techniques across the functional components of video content automation and detection systems. Each of the features was analysed in the light of implementation strategy, algorithms/models underlying, and datasets used for benchmarking and evaluating them. The following subsections outline the methodological categorisation applied to this study. The entire process flow of the proposed system is shown in Figure 1. The process starts with a user uploading a video, which is first passed through a preprocessing phase to normalise input and eliminate noise. The content is then scanned by the DeepFake detection module, which makes a decision on its authenticity. If the video is flagged as a DeepFake, it is sent for human validation to check and reject, if needed; otherwise, authentic content moves to the next step. For authentic videos, the image captioning module creates descriptive captions, which are enriched further by the hashtag recommendation module to propose context-aware and trending hashtags. The system then offers the created captions and hashtags to the user for validation. At this point, users can modify, refine, or cancel the system's proposals over and over again. Once the proposals are approved, the material is scheduled and then released on the website. This process not only guarantees the trustworthiness and authenticity of exposed media through DeepFake filtering but also increases user engagement through automated captioning and hashtagging, while maintaining user agency in the final decision-making process.

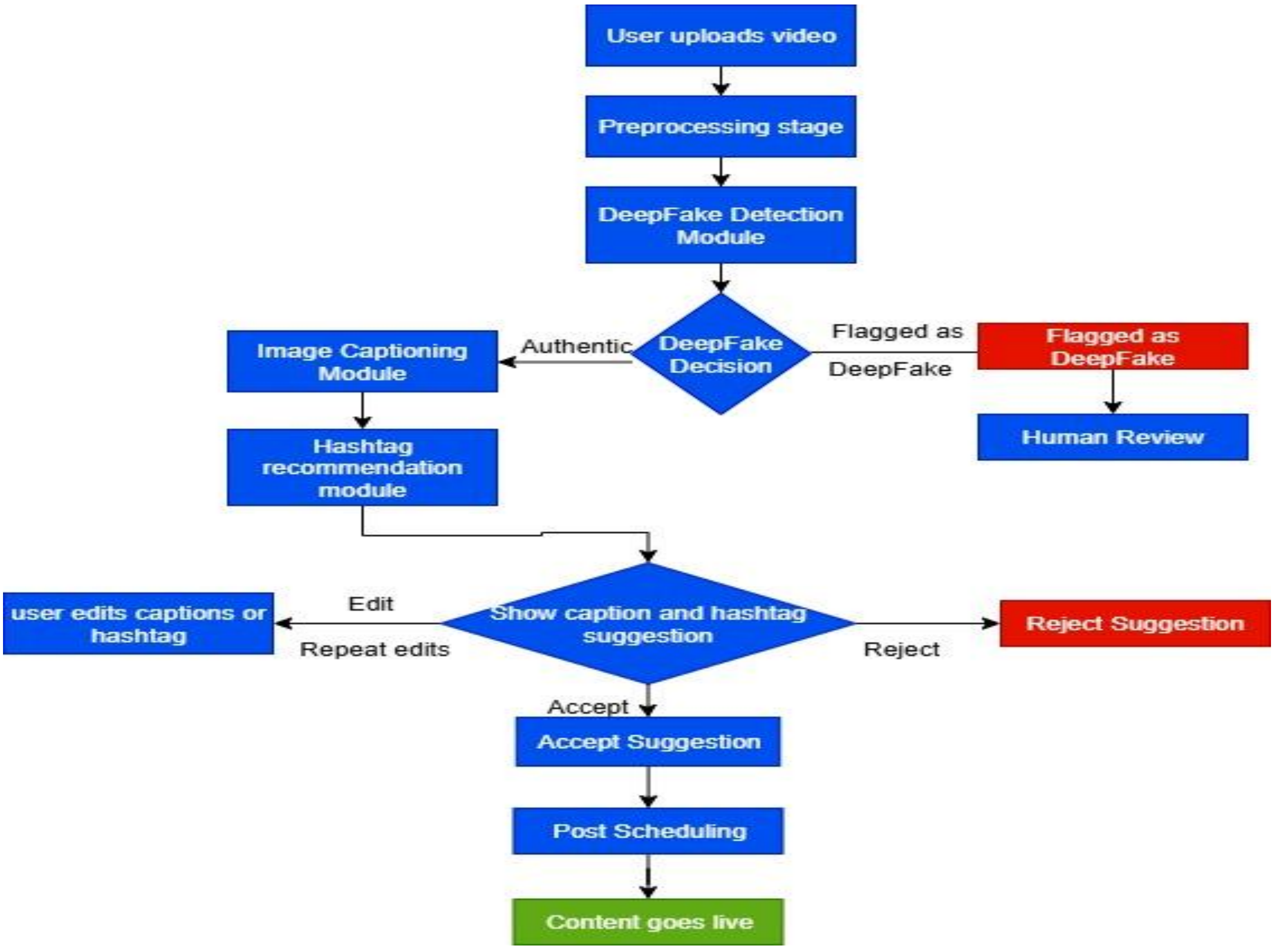

Figure 1: Workflow of proposed model

A. Preprocessing Stage

The preprocessing process in the proposed framework plays crucial to pre-process the uploaded video for the deepfake detection stage. As soon as the user uploads a video, it is segmented into separate frames so that frame by frame analysis is allowed. Face regions are identified and aligned from such frames in order to ensure uniformity in pose, scale, and position, thus minimizing variability due to light or camera viewpoints. Besides, preprocessing also entails eliminating noise and normalizing image characteristics like brightness and contrast so that unnecessary variations do not misguide the detection model. The frames are resized and transformed into the desired input format compatible with the deepfake detection network.

B. Deepfake Detection Module

The deepfake detection module starts following the preprocessing phase in which frames of videos are obtained for processing. This is done by using CNN based classifiers like MesoNet or a specialized CNN model. These models are trained on benchmark datasets such as FaceForensics++ and DFDC, allowing the system to determine if a video is real or fake. If a video has been detected as deepfake, it is forwarded for human validation, while valid content goes to the next pipeline stage. This ensures content reliability before publication.

C. Image Captioning Module

After a video is labeled authentic, it is processed into the image captioning module. This process makes use of transformer-based models such as BLIP or LXMERT to provide descriptive captions by drawing out visual features from video frames. The training datasets utilized are MS COCO and Conceptual Captions that offer high quality image text pairs to ensure precise caption generation. This module ensures that all uploaded videos are accompanied by an informative, contextually relevant caption.

D. Hashtag Recommendation Module

Following caption generation, the hashtag-suggestion module enhances content discoverability. As shown in Figure 1, it applies a keyword extraction method, namely KeyBERT with BERT embeddings, on the generated captions. Through correlating descriptive text with matching hashtags, the module increases personalization, search, and coverage of social media posts on different platforms.

IV. Future Directions

Although remarkable advancements have been realized in image captioning, hashtag suggestion, and detection of deepfakes, a number of research challenges persist, which provide avenues for future work.

- In the field of image captioning, most current models are based on large datasets with strong domain priors, restricting their ability to generalize to various real-world settings. Future work ought to explore domain adaptation and few shot learning techniques to provide stable captions in multiple contexts. Integrating multimodal transformers and vision-language pretraining models like CLIP and BLIP can further boost semantic richness and factual grounding of captions [1], [2]. Further, injecting user intent and personalization into captioning has not been explored much and can be a significant research area [3].
- In hashtag recommendation systems, existing methods tend to capture semantic content or user preferences, but less frequently both with comparable efficiency. Future platforms may use real-time trend detection using graph neural networks (GNNs) to dynamically adjust recommendations [4]. Likewise, low resource and multilingual hashtag generation is underexplored, and cross-lingual embedding spaces and zero shot
- learning approaches are needed to make it inclusive [5]. The other area with great potential is the application of ethical filters and misinformation detection to prevent hashtags from spreading toxic or misleading content [6].
- In deepfake detection, the biggest challenge is ensuring robustness to adversarially created and cross-platform manipulated content. Future work should aim at generalizable detection models trained on heterogeneous, in-the-wild datasets [7]. Multimodal fusion of audio, visual, and textual modalities, combined with

explainable detection paradigms, can enhance accuracy and transparency [8]. Additionally, the design of lightweight models that can be deployed to user devices can enable real-time security without cloud infrastructure dependency [9].

Last but not least, the broader context of AI-powered social media automation offers an opportunity to integrate these threads. A unified pipeline combining captioning, individualised hashtag suggestions, and strong verification before posting would both improve user interaction and ensure safer platforms. Designing such systems while addressing ethical, privacy, and policy considerations should be a key avenue for future work [10].

V. Conclusion

This review consolidates research on image captioning, hashtag suggestion, and deepfake detection, and examines their implications for AI-powered social media automation. In image captioning, the transition from CNN-RNN based models to attention-based and transformer-based models significantly enhanced semantic matching and descriptive precision [1]. Hashtag suggestion has evolved with end-to-end deep learning models with the ability to integrate textual, visual, and user-dependent inputs [2]. Deepfake detection has progressed towards multimodal methods and varied benchmarks, allowing for more robust comparative studies [3]. As a whole, these papers underscore the necessity of deep architectures, personalisation, and strong multimodality in guiding future AI systems. Even with remarkable progress, inherent limitations persist. Image captioning models are bound by dataset bias and poor transferability in unseen domains. Hashtag-suggestion approaches typically fail to reconcile semantic content extraction with flexible trend adaptation, thereby compromising relevance in rapidly evolving contexts. Similarly, while deepfake detection models have improved robustness, they still face challenges with generalisation across datasets and rapid advances in generative adversarial networks [4]. The literature remains fragmented when considering integration over tasks. Not many works attempt to bring captioning, hashtag personalisation, and verification pipelines together in a single social media automation setting. Low-resource and multilingual settings, where content generation and moderation are most challenging, lack a systematic focus. Ethical, explainable, and light-weight solutions are unexplored, hindering scale deployment. Upcoming research needs to adopt vision-language pre-trained transformers for captioning, graph-based trend modelling for hashtag recommendation, and generalizable multimodal frameworks for detecting deepfakes. Developing lightweight, real-time pipelines deployable on user-level devices, incorporating ethical and policy constraints, will be critical for the sustainability of automation. Overall, this review highlights progress but also identifies urgent gaps in generalisation, personalisation, and integration. By gathering dispersed research and identifying opportunities, this survey helps guide a research agenda toward reliable, effective, and ethically sound social media AI systems.

Reference

1. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 539-559.

2. Osman, A. A., Shalaby, M. A. W., Soliman, M. M., & Elsayed, K. M. (2023). A survey on attention-based models for image captioning. *International Journal of Advanced Computer Science and Applications*, *14*(2).
3. Chauhan, H. N., & Thacker, C. (2025). A comprehensive survey on automatic image captioning-deep learning techniques, datasets and evaluation parameters. *International Journal of Electrical and Computer Engineering (IJECE)*, *15*(3), 3257-3266.
4. Gupta, D., & Chakraverty, S. (2025). HaRNaT-A dynamic hashtag recommendation system using news. *Online Social Networks and Media*, *45*, 100294.
5. Djenouri, Y., Belhadi, A., Srivastava, G., & Lin, J. C. W. (2022). Deep learning based hashtag recommendation system for multimedia data. *Information Sciences*, *609*, 1506-1517.
6. Djenouri, Y., Belhadi, A., Srivastava, G., & Lin, J. C. W. (2022). Deep learning based hashtag recommendation system for multimedia data. *Information Sciences*, *609*, 1506-1517.
7. Połap, D. (2023). Hybrid image analysis model for hashtag recommendation through the use of deep learning methods. *Expert Systems with Applications*, *229*, 120566.
8. Bansal, S., Gowda, K., & Kumar, N. (2024). Multilingual personalized hashtag recommendation for low resource Indic languages using graph-based deep neural network. *Expert Systems with Applications*, *236*, 121188.
9. Mao, Q., Li, X., Liu, B., Guo, S., Hao, P., Li, J., & Wang, L. (2022). Attend and select: A segment selective transformer for microblog hashtag generation. *Knowledge-Based Systems*, *254*, 109581.
10. Ramanaharan, R., Guruge, D. B., & Agbinya, J. I. (2025). DeepFake video detection: Insights into model generalisation—A Systematic review. *Data and Information Management*, 100099.
11. Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, *252*, 124260.
12. Alrashoud, M. (2025). Deepfake video detection methods, approaches, and challenges. *Alexandria Engineering Journal*, *125*, 265-277.
13. Garg, D., & Gill, R. (2025). Unmasking Deepfakes: A Review of Current Datasets, Tools, and Detection Features. *Procedia Computer Science*, *259*, 1737-1748.
14. Sunil, R., Mer, P., Diwan, A., Mahadeva, R., & Sharma, A. (2025). Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. *Heliyon*, *11*(3).
15. Kumar, A., Singh, D., Jain, R., Jain, D. K., Gan, C., & Zhao, X. (2025). Advances in DeepFake detection algorithms: Exploring fusion techniques in single and multi-modal approach. *Information Fusion*, *118*, 102993.
16. Duhan, K., & Kajal, A. (2025). A Comparative Analysis of Deep Learning Based Approaches for DeepFake Identification. *Procedia Computer Science*, *259*, 482-493.
17. Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... & Etzioni, O. (2025). Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*.

18. Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, *16*, 100538.
19. Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Computer Law & Security Review*, *58*, 106162.
20. Kolosiuk, O. A., & Zinovatna, S. L. (2024). An automated social media manager based on artificial intelligence. *ІНФОРМАТИКА, КУЛЬТУРА, ТЕХНІКА Учредители: Odessa Polytechnic National University*, *1*(1), 124-132.