A Comprehensive Review of AI and Deep Learning Techniques for Detecting Cybersecurity Threats on Instagram

Mr. Dixitkumar M. Chaudhary[1], Dr. Idrish I. Sandhi[2]
[1] *Assistant Professor,* [2] *HOD-MCA,*
[1, 2]*Sankalchand Patel College of Engineering, Sankalchand Patel University, Gujarat, India*

Abstract: Instagram's focus on visuals—and the sheer number of people using it—has turned the platform into a magnet for phishing, impersonation, spam, fraud, and cyberbullying. Old-school keyword or rule-based detection just can't keep up, especially with the way people mix languages, use emojis, and hide bad behavior in images or weird text. This review pulls together research from 2016 to 2025 on how artificial intelligence and deep learning are changing the game when it comes to spotting cyber threats on Instagram. We dig into how threats show up on the platform, what makes Instagram unique, and how researchers build and label datasets. There's a lot of ground to cover—models that look at text, images, or both at once (think Transformers like BERT and mBERT, or neural nets like ResNet), tools that track relationships and timing, and ways to actually measure what works. We also look at explainable AI methods like SHAP, LIME, and Grad-CAM—because honestly, it helps to know why a model flagged something as dangerous. We put different methods side by side, share how they perform (F1, ROC-AUC, PR-AUC—you get the idea), and call out the big headaches that keep popping up: stuff like code-mixing, sarcasm, hidden abuse, sneaky URLs, adversarial attacks, privacy headaches, and biased datasets. Wrapping up, we lay out where Instagram-focused research should head next: stronger vision-language models, better network and time-based features, learning that happens right on your device or in a privacy-safe way, and smarter human moderation backed by dashboards you can actually understand.

Index Terms: Instagram, cybersecurity, deep learning, natural language processing, multimodal fusion, explainable AI, phishing, cyberbullying, impersonation, spam.

I. INTRODUCTION

You know, what's safe today might not be safe tomorrow. It kinda makes you think, doesn't it? Social media really changes how we connect and shop, but it also opens up new paths for people looking to stir up trouble. Social media has totally transformed how we talk, shop, and get our news all over the world, and it's been doing that for years. Instagram is totally a visual place now, all about creators. It really hooks people in with its core stuff: photos, those super short videos called Reels, temporary Stories, captions, hashtags, and all the influencer groups. What's cool is how all these elements help content spread quickly and get people talking. Meta's report says Instagram hit 3 billion active users each month by September 2025. This really shows how much Instagram has become a part of our daily online lives, both for culture and buying things. The sheer number of people on these platforms is a major part of why dealing with social media abuse is such a challenge. e.cybersecurity), it's like scammers are using pretty pictures and clever tricks to fool a bunch of folks without even trying that hard. It's still tricky to control content at a platform level. The rules about how to apply the law, how many folks are working on it, and how well machines actually work are always shifting. The quarterly data they release really shows how much things are changing, trying to balance everyone's right to "free expression" with reducing harm. WIRED+ 5Reuters +5The Verge+5.

Unlike text-focused platforms, Instagram's multimodal setup lets attackers weave together image cues—the shimmer of a logo, for instance—with text

and network signals. You'll often see brand knockoffs that mimic logos and fonts, fake profiles that steal identities, and sneaky "link in bio" traps meant to snatch logins. Then come the money schemes—giveaway frauds, crypto-doubling promises—plus waves of coordinated spam, engagement fakery, and harassment that hides behind memes and sharp, taunting captions. These behaviors hide within design cues like logos, packaging, or templates, flare up in brief story cycles, and move through social webs of followers and mentions—so single-modality detectors stumble, since the real meaning lies in how signals mix and ripple across channels. Research groups are now pulling this challenge together under multimodal content, like blending image and sound data into a single stream. Moderation, with dedicated workshops and benchmarks highlighting the need for platform-aware methods 777. Multimodal Content Moderation

Recent breakthroughs in AI and deep learning provide very capable components for Instagram, specific detection. Vision backbones (CNNs and more recently vision transformers) can identify brand logos, recognize visual scams, and detect tampered or synthetic media. Language models (Transformers/LLMs) are really good at understanding captions, hashtags, and DMs at a very nuanced level, and multimodal LLMs (MLLMs/LVLMs) align image, text features to reason over posts in a holistic manner. At the same time, safety for multimodal models has become a separate area, with documentation of attack surfaces, evaluation protocols, and defenses relevant to trust & safety deployments 888999.

Moreover, graph neural networks (GNNs) are used beyond per, post modeling to capture the relational and temporal structure follower ties, interaction motifs, and diffusion paths which have been demonstrated as efficient for fraud, misinformation, and coordinated, behavior detection in social systems 101010111111.SpringerLink+3IJCAI+3ACM Digital Library+3

Looking at it from an operational standpoint, handling moderation for an Instagram, size platform simply can't be done by humans alone; it needs automation complemented by humans. The platform's transparency reports reveal that the company has been continuously reviewing its precision/recall trade, offs, trying to lower the number of false, positive enforcement, and changing the policy taxonomies that guide the enforcement actions across Facebook, Instagram, Messenger, and Threads. 444, 666, 121212. For researchers, this motivates not just higher raw accuracy but also robustness (to adversarial adaptation, multilingual variation, and distribution shift), explainability (to support reviewer decisions and appeals), and privacy-preserving learning (given sensitive user data). Facebook+2WIRED+2

Scope and goals. This paper provides a platform-specific, comprehensive review of AI and deep learning techniques for detecting cybersecurity threats on Instagram and closely related visual-first ecosystems. We: (i) propose an Instagram-oriented threat taxonomy spanning phishing, impersonation, spam/fraud, coordinated inauthentic behavior, and cyberbullying; (ii) survey datasets, collection pipelines, and labeling practices, noting gaps in multimodal annotations and graph/temporal ground truth; (iii) analyze text-only, vision-only, multimodal, and graph-temporal detection approaches, including pretraining, cross-modal fusion, and weak supervision; (iv) summarize evaluation protocols (metrics, robustness checks, cross-domain transfer) and explainability frameworks suitable for moderation workflows; and (v) discuss open challenges—data scarcity, adversarial robustness, long-horizon temporal reasoning, privacy, and XAI for reviewer-in-the-loop operations. Collectively, our goal is to consolidate actionable insights for researchers, practitioners, and policy stakeholders to build context-aware, resilient, and auditable detection systems for Instagram.

II. SCOPE AND METHOD

Instagram and other visual text based social networks such as Facebook and Twitter and TikTok. In terms of issue coverage we look at phishing, fake accounts, spam, commercial abuse, and cyberbullying. We only included in our review papers which had (a) new algorithms, (b) that did empirical study using common measures of success, and (c) which look at the Instagram environment which has visual and textual

content. Also we gave preference to studies which made their data open or which present code that can be reproduced.

The review organizes research by:

- Data Modality: Text based (captions, comments), image based (posts, profiles) and multichannel (text image).
- Auxiliary Information: Network based (user connections) and temporal (activity over time).

Data from Google Scholar, IEEE Xplore, ScienceDirect, ACM Digital Library, SpringerLink and arXiv was used which included the terms "Instagram", "cybersecurity", "AI detection" and "multimodal learning".

In each paper we looked at issue definition, dataset use, model design, features presented, evaluation methods, and results which in turn enabled a structured comparison and we identified key research issues in AI based Instagram threat detection.

III. THREAT TAXONOMY AND PLATFORM CHARACTERISTICS

We use a five-part taxonomy common on Instagram—phishing, impersonation, spam, cyberbullying or harassment, and benign—like sorting posts into neat drawers after a busy day online. Phishing often uses catchy captions, link-in-bio tricks, and shortened URLs, along with screenshots of fake login or reward pages tucked into images that look perfectly legitimate. Impersonation happens when fake accounts copy public figures or brands, using look-alike usernames, matching profile photos, and posts that echo the same tone. Spam can be anything from endless promo posts to comment floods packed with loud, pushy hashtags. Cyberbullying shows up as insults, threats, and planned attacks—messages dripping with sarcasm, packed with slang, local phrases, and bursts of bright emojis. Instagram's own tools—things like overlays, stories, reels, stickers, and hashtags—shape the way each threat shows up, as if the platform itself colors the danger with its bright filters and quick scroll.

Table 1. Instagram Threat Taxonomy and Modality Cues

Threat Type	Primary Cues	Modalities	Notes
Phishing	URLs, call-to-action phrases, credential requests	Text+Image	Often uses screenshots of forms or brand logos.
Impersonation	Username similarity, profile photo, brand assets	Image+Metadata	Facial/logo similarity; verification badge absence.
Spam	Repetitive hashtags, promo templates, bulk posting	Text+Image	Bot-like behavior; link shorteners.
Cyberbullying	Insults, slurs, sarcasm, harmful emojis	Text	Context dependence; code-mixing and implicit abuse.
Benign	No malicious cues	Any	Potential false positives near promos or satire.

IV. DATASETS AND ANNOTATION PRACTICES

The main obstacles remain high-quality datasets. Due to Instagram's terms and privacy policies, sharing data is restricted, resulting in many studies having to gather only publicly available data or using proxy platforms (e.g., Twitter) to shift data and methods. When datasets from Instagram become available, they frequently do not contain text–image pairs, contain inadequate multilingual representation and lacking balance in coverage of the classes. When the annotators are trained, inter-annotator agreement (e.g., Cohen's Kappa) is often the only metric of reliability reported, and class imbalance is dealt with by using stratified sampling, reweighting, or augmentation. Suffice it to say, the data has to be used responsibly and in compliance with the policies of the platforms.

Some labeling challenges are ambiguity (e.g. satire vs. spam), culturally specific insults, and unclear phishing intent without links. Labeling tools like Label Studio automate annotations. For spam, fraud, and phishing, quality control works best with gold standard, dual annotation, and adjudication. For multilingual content, the preservation of code-switch, transliteration, and addressed semantic layering in annotation tokenization adjust the span of content multilingualism.

V. TEXT-ONLY MODELS (NLP)

Transformer encoders (BERT, RoBERTa, mBERT, XLM-R) dominate text classification on social media. Fine-tuning on captions and comments yields strong baselines, and sequence models such as Bi-LSTMs stacked over contextual embeddings can capture conversational flow across comments. Preprocessing steps include placeholder tokens for URLs/mentions, emoji normalization, and length constraints (e.g., 128–256 tokens). For multilingual content, mBERT/XLM-R and lightweight adapters or LoRA are useful. Classical baselines (TF–IDF + SVM, logistic regression) remain competitive in resource-limited settings and as interpretable yardsticks.

- URL and handle masking to reduce overfitting to specific strings.
- Hashtag segmentation to recover words from CamelCase (e.g., #FreeGiftCard).
- Class-weighted losses and focal loss to address minority classes.
- Data augmentation via back-translation and synonym substitution (with care for toxicity tasks).

VI. IMAGE-ONLY MODELS (COMPUTER VISION)

CNN backbones (ResNet, EfficientNet) and, increasingly, vision transformers (ViT) detect visual cues such as brand logos, text overlays, manipulated screenshots, nudges like arrows and buttons, and stylistic patterns of spam templates. Transfer learning from ImageNet is standard; augmentations (rotation, blur, color jitter) improve robustness to user edits and compression artifacts. OCR features extracted from images can be fused back into NLP pipelines to catch text embedded in images (e.g., phishing CTAs).

VII. MULTIMODAL FUSION AND METADATA

Fusion approaches combine caption/comment embeddings with visual features and, optionally, user/profile metadata. Late fusion concatenates modality representations before classification; early fusion aligns token and patch embeddings; joint fusion leverages cross-attention (e.g., CLIP-like or BLIP-style objectives). Multimodal models consistently outperform single-modality baselines on Instagram-like tasks due to complementary signals. Graph features (follower–following networks, interaction graphs) and temporal dynamics (posting intervals, burstiness) further boost performance for spam and impersonation detection.

- Late fusion: simple and robust; easy to ablate and explain.
- Early/joint fusion: stronger but requires careful alignment and more data.
- Graph neural networks (GNNs): propagate signals across user/content networks.
- Temporal models: detect anomalies via inter-post timing and seasonality.

Table 2. Representative Modeling Choices and Trade-offs

Modality	Representative Methods	Pros	Cons
Text	BERT/mBERT, XLM-R, Bi-LSTM, TF–IDF+SVM	Strong baselines; multilingual support	Implicit abuse and sarcasm remain hard
Image	ResNet, EfficientNet, ViT, OCR	Captures logos/templates; works without text	Fails on text-only attacks; data hungry
Multimodal	Concatenation (late), cross-attention (joint), CLIP/BLIP	Best accuracy; complementary cues	Complex training; alignment challenges
Graph/Temporal	GNNs, anomaly detection, Hawkes processes	Captures behavior patterns and botnets	Data access/privacy constraints

VIII. EXPLAINABLE AI (XAI) FOR MODERATION

Explainability supports trust, auditing, and appeals in moderation. SHAP provides global and local feature attributions for text and fused representations; LIME builds local surrogate models for case-level explanations; Grad-CAM highlights salient image regions influencing CNN decisions. Integrated Gradients and attention rollout offer alternatives for transformer-based models. Dashboards that display saliency maps, top textual cues, and system confidence help reviewers make faster, fairer decisions and identify systematic errors.

IX. EVALUATION PRACTICES AND METRICS

Robust evaluation reports per-class precision, recall, and F1 (macro/micro), ROC-AUC and PR-AUC (especially under class imbalance), and confusion matrices. Cross-validation with stratified folds and a held-out test set is common. Ablation studies quantify each modality's contribution; error analyses examine failure cases like sarcasm and obfuscated URLs. Operational metrics such as latency, throughput, and moderation workload reduction are important for deployment realism.

- Use macro-F1 to avoid dominance by benign class.
- Calibration metrics (ECE) for thresholding and human handoff.
- Fairness slices (language, region) to detect disparate impact.
- Adversarial robustness checks (typos, homoglyphs, image perturbations).

X. COMPARATIVE SYNTHESIS OF METHODS

Across studies, multimodal fusion typically surpasses text-only or image-only baselines, with reported accuracy often improving by 5–10 absolute points depending on dataset composition. Text-only transformers remain competitive for cyberbullying when imagery provides little signal, while image channels are crucial for phishing with screenshot overlays and brand spoofing. Graph-temporal features are particularly effective for impersonation and spam where behavior signals dominate content semantics.

Table 3. High-level Comparison Across Common Instagram Threat Tasks

Study Focus	Modality	Backbone	Target Threat(s)	Typical Metrics (Reported Ranges)
Cyberbullying/Toxicity	Text	BERT/mBERT, XLM-R	Harassment/abuse	Macro-F1 ≈ 0.80–0.90; ROC-AUC ≈ 0.85–0.95
Phishing/Scam	Text+Image	ResNet + Transformer	Phishing, scams	Accuracy ≈ 0.85–0.93; Macro-F1 ≈ 0.84–0.92
Impersonation/Fake Accounts	Multimodal+Graph	CNN/ViT + GNN	Impersonation	F1 ≈ 0.85–0.92; AUC ≈ 0.90–0.96
Spam/Promo Abuse	Text/Graph	Classical + Transformer	Spam	F1 ≈ 0.82–0.90; PR-AUC high at high recall

XI. CHALLENGES AND OPEN PROBLEMS

- Multilingual and code-switched content with slang, dialects, and transliteration.
- Implicit abuse, sarcasm, and context dependence requiring discourse understanding.
- URL obfuscation and link-in-bio strategies; OCR quality on compressed screenshots.
- Adversarial attacks (token perturbations, image masking) and model robustness.
- Data scarcity and sharing restrictions; privacy-preserving data collection and annotation.
- Bias and fairness across languages, cultures, and demographics; transparency and appeals.
- Real-time constraints on-device or near-real-time moderation with limited compute.

XII. RESEARCH ROADMAP FOR INSTAGRAM

- Foundation vision-language models (e.g., CLIP-like, BLIP-style) adapted for Instagram with contrastive pretraining on image–caption pairs.
- Unified multimodal transformers with cross-attention to jointly encode text, image, and metadata; adapters/LoRA for efficient fine-tuning.
- Graph-temporal integration: user/content GNNs, session-level models, and anomaly detection pipelines.
- Privacy-preserving learning (federated, on-device) and secure aggregation; differential privacy for analytics.

- Model compression and distillation for mobile/edge deployment; mixed precision and early-exit strategies.
- Human-in-the-loop workflows with calibrated thresholds, XAI dashboards, and redress mechanisms.
- Robustness evaluation against adversarial and obfuscation tactics; watermark/OCR improvements for image text.

XIII. PRACTICAL DESIGN GUIDELINES

1. Start with strong text and image baselines (BERT/mBERT + ResNet/EfficientNet); evaluate independently before fusion.
2. Engineer preprocessing for Instagram specifics: hashtag segmentation, emoji handling, URL placeholders, OCR extraction.
3. Prefer late fusion for simplicity and ablations; evolve to cross-attention/joint fusion as data grows.
4. Leverage class-weighted/focal losses and macro-F1 optimization; include calibration for decision thresholds.
5. Integrate SHAP/LIME/Grad-CAM in reviewer UI; log explanations for auditing and model refinement.
6. Track behavior signals (posting cadence, follower dynamics) under privacy guardrails; consider GNNs.
7. Plan for deployment: inference time budgets, batching, and fail-safe human escalation policies.

XIV. CONCLUSION

Instagram presents a uniquely multimodal and adversarial environment for cybersecurity threat detection. The literature consistently shows that multimodal fusion—augmented with graph-temporal signals and explainability—delivers the most reliable performance across phishing, impersonation, spam, and cyberbullying tasks. The next wave of research and deployment will hinge on sharper vision-language models, privacy-safe training, and moderation that keeps people at the center—like a watchful hand guiding the work. This review pulls together what we know so far and lays out a clear roadmap shaped by how Instagram actually works—the fast scroll, the constant buzz of new posts.

ACKNOWLEDGMENT

I'm thanks to Shri Prakashbhai Patel, Dr. P. M. Udani, Dr. P. J. Patel and Dr. M. I. Sandhi from Sankalchand Patel University for their support.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016.
[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, 2017.
[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. KDD, 2016.
[5] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Comput. Surveys, vol. 51, no. 4, 2018.
[6] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Social Media: A Comparative Study," Inf. Process. & Management, 2020.
[7] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. ICML, 2021. (example for ViT/DeiT)
[8] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv:2103.00020, 2021. (CLIP)
[9] Y. Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv:2301.12597, 2023.
[10] G. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in Proc. ICLR, 2017.
[11] T. Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," in Proc. ICML, 2020.
[12] S. Ruder et al., "A Survey of Cross-lingual Transfer Learning in NLP," JAIR, 2021.
[13] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017.
[14] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution

Examples in Neural Networks," in Proc. ICLR, 2017. (Calibration)

[15] D. Dua and C. Graff, "UCI Machine Learning Repository," 2019. (as a generic dataset repository reference)

[16] D. M. Chaudhary and I. I. Sandhi, "AI-Powered Detection and Classification of Cybersecurity Threats on Instagram Using Deep Learning and NLP," 2025. (Reviewed work)