Tetiana Cheberko. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 15, Issue 11, November 2025, pp 83-99

RESEARCH ARTICLE OPEN ACCESS

AI-Driven Bot Detection and Mitigation in Social Apps: A Case Study of SOYO

Tetiana Cheberko
University of Florida, Warrington College of Business, USA

Abstract
The rapid growth of social platforms in emerging markets has created new opportunities for communication and community building, but it has also introduced new risks. Automated chatbots that impersonate real users have become increasingly common, often exploiting platform vulnerabilities to target users with social engineering schemes and fraudulent demands. SOYO, a major social application in Kenya and Nigeria, faced significant disruptions when automated bots began generating large volumes of deceptive messages. This study presents an in-depth analysis of SOYO's hybrid artificial intelligence system designed to detect, validate, and mitigate bot activity at scale. The system integrates natural language processing for message-level analysis, behavioral anomaly detection for real-time monitoring of user actions, and convolutional neural network-based face verification for identity confirmation. Data from two million users over a three-month period were examined to evaluate the performance of the system. The results show a 96 percent decline in bot-generated messages and an 87 percent reduction in user complaints. This paper details the architectural design, model selection, operational outcomes, and broader implications for AI-driven platform security in emerging markets.
Keywords: AI, social platforms, bot detection, NLP, CNN, behavioral anomaly detection, SOYO

Date of Submission: 05-11-2025 Date of acceptance: 16-11-2025

I. Introduction

Social networking platforms in emerging digital markets face unique challenges due to high user engagement, rapid adoption, and relatively limited regulatory oversight. These conditions create an environment highly susceptible to automated malicious actors, including bots, fraudulent accounts, and social engineering scripts. SOYO, a social application that rapidly became one of the top-ranked platforms in Kenya and Nigeria, experienced a dramatic increase in such malicious activity as its user base grew.

The proliferation of automated accounts posed severe risks to the platform's integrity, user safety, and long-term trust. Bots sent unsolicited messages, often requesting personal information, financial contributions, or performing repetitive, scripted interactions that mimicked human users. Prior to implementing AI-based defenses, SOYO's moderation team observed that nearly every support ticket was related to bot activity, with complaint rates reaching 100% of daily user reports. The exponential growth of bot accounts quickly overwhelmed human moderation, creating an urgent need for automated detection and mitigation solutions.

The challenges faced by SOYO are consistent with global trends in social media. Automated bots can replicate human-like behaviors, utilize generative AI to craft realistic messages, and engage in multi-account coordination. Traditional rule-based filters and manual moderation methods are insufficient in such dynamic environments. To address this, platforms must employ multi-layered AI-driven defenses that integrate behavioral analysis, content inspection, and identity verification.

SOYO's approach combined several AI methodologies to detect and prevent fraudulent activity effectively. First, **natural language processing (NLP)** algorithms were applied to incoming messages to identify requests for personal information, financial solicitations, or other patterns indicative of malicious intent. These NLP models were trained using a combination of supervised learning on labeled historical chat data and semi-supervised techniques to adapt to emerging patterns. Second, **behavioral anomaly detection** was employed to monitor user activity metrics, such as session duration, frequency of message sending, and repetition of identical content. Accounts exhibiting statistically abnormal behaviors were flagged for review or automatic intervention. Third, a **deep-learning-based face verification system**, leveraging convolutional neural networks (CNNs), ensured that each account corresponded to a real human user. Suspicious accounts failing verification

were prompted to undergo re-verification before resuming platform access.

The implementation of this hybrid system yielded substantial improvements in platform security. Following deployment, SOYO achieved a **96% reduction in bot-generated chats**, and the rate of user complaints dropped from **100% to 13%**. These metrics underscore the effectiveness of combining content analysis, behavioral modeling, and biometric verification to mitigate automated attacks in social applications. Beyond quantitative improvements, the system also enhanced user confidence, retention, and overall platform credibility.

This case study provides valuable insights into the operationalization of AI for security in emerging market social apps. By documenting the challenges, AI strategies, and measurable outcomes, this paper contributes to the growing literature on bot detection, automated fraud prevention, and safe platform design. It demonstrates that scalable AI systems can reduce reliance on human moderation, adapt to evolving threats, and maintain the authenticity of user networks.

The remaining sections of this article will provide a detailed methodology of the AI models, describe the results with evaluation metrics, discuss the implications and limitations of the system, and conclude with lessons learned and potential directions for future research in AI-driven platform security.

II. Methods

To address the escalating issue of bot accounts and fraudulent activity on SOYO, the platform implemented a **hybrid artificial intelligence (AI) system** that integrated three primary modules: (1) natural language processing (NLP) for message content analysis, (2) behavioral anomaly detection to monitor user activity patterns, and (3) convolutional neural network (CNN)–based face verification to ensure account authenticity. This multi-layered approach enabled the platform to detect and mitigate malicious activity in real time, reducing reliance on human moderation while maintaining a high level of accuracy.

2.1. NLP-Based Message Content Analysis

The first line of defense focused on analyzing the content of user messages. Malicious actors commonly send messages attempting to elicit personal information or solicit financial transactions. SOYO utilized a supervised NLP classifier trained on a corpus of over 1 million chat messages, labeled for spam, scam, and benign content. The classifier leveraged **transformer-based models**, similar to BERT (Bidirectional Encoder Representations from Transformers), to capture contextual relationships between words and detect subtle semantic patterns indicative of malicious intent.

The preprocessing pipeline included tokenization, stop-word removal, and normalization of text to handle slang, abbreviations, and multilingual content common in Kenya and Nigeria. Messages were encoded as embeddings and passed through the neural network for classification. The model produced a probability score for each message indicating the likelihood of being malicious. Thresholds were set based on precision-recall tradeoffs to minimize false positives while effectively capturing spam and scam messages. Suspicious messages were blocked automatically, while borderline cases were flagged for further review or user verification.

2.2. Behavioral Anomaly Detection

While message content provides valuable information, sophisticated bots often adapt their language to avoid detection. Therefore, SOYO incorporated a behavioral anomaly detection module to analyze patterns of user activity. Key features included session duration, message frequency, intervals between messages, repetition of identical content, and temporal patterns across multiple accounts.

Data were collected continuously and represented as time-series vectors, which were then processed using **unsupervised machine learning techniques**, including clustering (DBSCAN) and autoencoder-based anomaly detection. Accounts exhibiting statistically unusual behaviors were assigned anomaly scores. High-scoring accounts triggered automated interventions, such as temporary message restrictions or mandatory re-verification. Behavioral rules were continuously updated to incorporate newly observed bot strategies, ensuring the system remained adaptive to evolving threats.

The behavioral module also included cross-account correlation. Bots often operate in groups, sending identical messages simultaneously. By calculating similarity matrices across accounts and clustering coordinated behaviors, the system could detect and mitigate coordinated attacks.

2.3. CNN-Based Face Verification

To guarantee that accounts represented real human users, SOYO integrated a **face verification system** based on convolutional neural networks (CNNs). During onboarding, users were prompted to submit a profile photo. The CNN was trained on a large dataset of labeled facial images to extract discriminative facial embeddings. At login or when suspicious behavior was detected, users were

required to provide a live selfie for verification. The embeddings were compared with the registered image using cosine similarity metrics.

Accounts failing verification were temporarily restricted, and users were prompted to retry verification. This process ensured that bots or fake accounts could not persist on the platform. The CNN system was optimized for speed and robustness, capable of verifying users in under two seconds while maintaining high accuracy across different lighting conditions, ethnicities, and device cameras.

2.4. System Architecture and Integration

The three AI modules were integrated into a **centralized risk scoring engine**, which calculated an overall risk score for each account based on message content, behavioral anomalies, and face verification outcomes. Scores above a predetermined threshold resulted in automated mitigation actions, including message blocking, temporary suspension, or forced re-verification.

The architecture supported **real-time processing**, using stream-based data pipelines built on Apache Kafka for message ingestion and TensorFlow Serving for deploying the NLP and CNN models. Behavioral anomaly detection models were updated daily using a batch processing framework to incorporate the latest activity data.

2.5. Threshold Selection and Performance Tuning

Thresholds for classification and anomaly detection were determined empirically through cross-validation on historical data. The NLP classifier threshold was optimized to achieve a precision of 95% and recall of 93%, balancing false positives and false negatives. Behavioral anomaly detection thresholds were set to capture 90% of known bot accounts while minimizing disruption to genuine users. Face verification similarity thresholds were tuned to achieve a false acceptance rate below 1% and a false rejection rate below 5%.

2.6. Continuous Learning and Adaptation

Recognizing that bot strategies evolve rapidly, SOYO implemented continuous learning mechanisms. Misclassified messages and accounts flagged by users were incorporated into retraining datasets. Periodic retraining ensured that both NLP and behavioral models adapted to new attack vectors, slang, and messaging patterns.

2.7. Security and Privacy Considerations

User privacy was a critical concern. Personal data collected for verification and behavioral analysis were encrypted in transit and at rest. Facial images used for CNN verification were stored temporarily and discarded after verification or re-verification. All AI models complied with local data protection regulations and the platform's privacy policies, ensuring ethical use of sensitive information.

Summary of Methods

Through the integration of NLP content analysis, behavioral anomaly detection, and CNN-based face verification, SOYO created a **robust, adaptive, and real-time AI system**. This hybrid approach allowed the platform to identify and mitigate bot accounts and fraudulent activity effectively, improving user safety, reducing complaints, and maintaining platform integrity.

III. Results

The deployment of the hybrid AI system on SOYO yielded significant improvements in detecting and mitigating bot activity, fraudulent messaging, and unauthorized account creation. The evaluation of system performance was conducted over a **three-month period following deployment**, encompassing over **2 million active users** across Kenya and Nigeria. The analysis focused on three primary dimensions: reduction in bot-generated chats, user complaints, and accuracy of detection modules.

3.1. Reduction in Bot-Generated Chats

Prior to AI implementation, daily bot-generated messages accounted for approximately 60% of all message traffic, severely degrading user experience. Following deployment, the **NLP-based message filtering** system successfully intercepted 96% of malicious messages before delivery. The classifier effectively identified attempts to extract personal information, solicit money, or distribute repetitive spam.

Table 1 summarizes the reduction in bot message volume:

Metric	Pre-AI Deployment	Post-AI Deployment	Change (%)
Bot-generated messages/day	1,200,000	48,000	-96
User-reported bot interactions	5,500	720	-87
Spam complaints	100% of reports	13% of reports	-87

The precision of the NLP classifier was measured at **95%**, with a recall of **93%**, demonstrating a robust ability to capture malicious content while minimizing disruption to genuine users. The high accuracy contributed directly to the drastic reduction in user-reported spam complaints.

3.2. Behavioral Anomaly Detection Performance

The behavioral anomaly detection module evaluated users' activity patterns, including message frequency, session duration, and repetitive messaging behaviors. Among 50,000 flagged accounts during the evaluation period, **92% were confirmed to be bots**, while the remaining 8% included some highly active genuine users who required additional verification.

Figure 1 (placeholder) illustrates the distribution of anomaly scores before and after AI deployment. Users with high anomaly scores underwent re-verification, effectively preventing bot accounts from re-entering the platform. The module also detected coordinated bot campaigns that manually curated content or attempted to bypass NLP filters, demonstrating the value of behavioral analytics as a complementary defense mechanism.

3.3. Face Verification Outcomes

The CNN-based face verification system validated the authenticity of user accounts. During the evaluation period, approximately **150,000 users** underwent verification due to suspicious behavior or new account registration. The system achieved:

- **False Acceptance Rate (FAR):** 0.8%
- **False Rejection Rate (FRR):** 4.5%

This high accuracy ensured that the majority of active accounts corresponded to real humans. Suspicious accounts failing verification were temporarily suspended, preventing fraudulent interactions and preserving user trust.

3.4. Overall Platform Impact

The integrated AI system significantly improved overall platform safety and user satisfaction. Key outcomes include:

- **Reduction in user complaints:** Prior to AI implementation, 100% of daily support tickets were related to bot or spam activity. Post-deployment, complaints dropped to 13%, representing an **87% improvement**.
- **Retention of genuine users:** By automatically blocking only malicious accounts and minimizing false positives, genuine users experienced fewer interruptions, which supported **higher retention rates**.
- **Operational efficiency:** Manual moderation efforts decreased substantially, allowing the moderation team to focus on complex cases rather than routine spam or bot activity.

3.5. Module Interactions and System Synergy

The combined use of NLP, behavioral analysis, and face verification created a synergistic effect. NLP filters intercepted content-level threats, behavioral monitoring captured subtle or coordinated bot activity, and face verification ensured human authenticity. This multi-layered approach prevented bots from exploiting any single system weakness.

During the evaluation period, **no large-scale bot breaches occurred**, indicating the robustness of the integrated system. Additionally, the AI modules provided **real-time alerts** for emerging threats, allowing for proactive intervention.

3.6. Statistical Evaluation

To quantitatively assess performance, standard evaluation metrics were calculated:

- **Precision:** 94% overall (weighted across modules)
- **Recall:** 92% overall
- **F1 Score:** 0.93
- **Reduction in bot-related interactions:** 96%
- **Decrease in user complaints:** 87%

These results indicate that the system not only efficiently identified bots but also minimized disruption to legitimate users. Comparative analysis with pre-AI performance showed statistically significant improvements ($p < 0.01$) in all metrics.

3.7. Case Examples

Several specific incidents illustrate system effectiveness:

1. A coordinated bot network attempted to distribute phishing messages to 3,000 users simultaneously. The behavioral anomaly detection module identified repetitive patterns within seconds, blocked all accounts, and initiated re-verification for remaining suspicious users.
2. A new bot type used conversational scripts to imitate genuine users, bypassing the initial NLP filter. The combination of behavioral scoring and face verification successfully intercepted these accounts, preventing large-scale exposure.

3.8. Limitations of Results

While the system demonstrated remarkable improvements, certain limitations exist. Very sophisticated bots using AI-generated images for profile pictures occasionally bypassed initial verification, requiring additional monitoring. Some highly active human users triggered false positives, highlighting the importance of continuous threshold tuning.

IV. Discussion

The implementation of the hybrid AI system on SOYO provides critical insights into the practical application of artificial intelligence in social platforms operating in emerging markets. By integrating **NLP-based content analysis, behavioral anomaly detection, and CNN-based face verification**, SOYO achieved measurable reductions in bot-generated messages and user complaints. This multi-layered approach demonstrates both the effectiveness and adaptability of AI-driven security interventions in dynamic online environments.

4.1. Significance of AI Integration

The results indicate that combining multiple AI techniques significantly enhances the reliability of bot detection. While single-method approaches—such as message filtering or behavior monitoring alone—can detect specific types of malicious activity, they are often insufficient against sophisticated bot strategies. Bots today can mimic human conversational patterns, coordinate activity across multiple accounts, and even generate AI-driven images to simulate real users.

SOYO's **integrated architecture** successfully mitigated these threats. NLP-based filters intercepted messages containing personal data requests or financial solicitations, while behavioral anomaly detection captured accounts exhibiting non-human activity patterns. Face verification ensured that automated accounts could not maintain persistent access. The synergy between these modules created a **robust defense** that is resilient against evolving attack vectors.

4.2. Comparison with Existing Literature

Previous studies have highlighted the challenges of bot detection in social networks. For instance, Cresci et al. (2017) noted that coordinated botnets are capable of bypassing rule-based filters, necessitating behavioral and content analysis combined with machine learning for effective detection. SOYO's results align with these findings, demonstrating that a hybrid AI approach can achieve high precision and recall in live, large-scale environments.

Additionally, research by Zhang et al. (2020) emphasized the importance of incorporating identity verification into bot mitigation strategies. The use of CNN-based face verification on SOYO reinforces this recommendation, ensuring that detected anomalies are indeed associated with human or non-human accounts. The empirical metrics—96% reduction in bot messages, 87% decrease in user complaints—indicate that combining content, behavior, and biometric verification significantly outperforms single-method interventions.

4.3. Operational Implications

The deployment of AI in SOYO's ecosystem yielded several operational advantages:

1. **Reduction in Manual Moderation:** Prior to AI deployment, the moderation team spent significant resources responding to spam reports and verifying accounts. Automation allowed the team to focus on exceptional cases requiring human judgment.
2. **Real-Time Mitigation:** The AI modules processed user activity and messages in real time, enabling immediate intervention. This immediacy is critical for preventing rapid dissemination of malicious content in social platforms.
3. **Adaptability:** Continuous learning mechanisms ensured that NLP and behavioral models adapted to new attack strategies, slang, and emerging messaging patterns. This dynamic adaptability is essential in environments where malicious actors continuously evolve.

4.4. Impact on User Experience and Retention

By significantly reducing bot interactions, SOYO improved user trust and engagement. Users reported fewer unsolicited messages, lower exposure to scams, and increased confidence in account authenticity. The precise targeting of AI interventions minimized disruption to legitimate users, which is crucial for retention.

These outcomes are particularly relevant in emerging markets, where platforms must compete for attention while maintaining credibility and safety. The measurable improvement—from 100% user complaints related to bots to just 13% post-deployment—illustrates the positive influence of AI on user satisfaction and overall platform stability.

4.5. Scalability and System Performance

The AI system's architecture, built with real-time streaming pipelines (Apache Kafka) and TensorFlow Serving for model deployment, demonstrates scalability. As the user base grows, modules can be parallelized or distributed across cloud resources to maintain low latency and high throughput.

Behavioral anomaly detection scales efficiently, as time-series and clustering analyses can process large volumes of user activity data. CNN-based verification, while computationally intensive, was optimized for speed, ensuring minimal user friction during verification. Scalability considerations also include model retraining frequency and the volume of flagged accounts; SOYO adopted a hybrid approach combining batch

retraining with continuous monitoring for adaptive learning.

4.6. Limitations and Challenges

Despite the successful outcomes, several limitations warrant discussion:

- **False Positives:** Highly active genuine users occasionally triggered alerts, requiring additional verification. Continuous threshold tuning and manual review of borderline cases are essential to mitigate negative user experiences.
- **Evolving Bot Strategies:** Sophisticated bots using AI-generated images or advanced conversation models could potentially bypass initial detection. Ongoing research and periodic model updates are necessary to maintain system effectiveness.
- **Privacy Concerns:** While facial verification improved security, collecting biometric data raises privacy considerations. SOYO implemented encryption and temporary storage policies, but compliance with regional regulations and user consent remains critical.

4.7. Lessons for Emerging Market Platforms

SOYO's experience offers several transferable lessons:

1. **Hybrid AI is More Effective:** Combining multiple AI modalities (content, behavior, identity) is more robust than single-method approaches.
2. **Continuous Learning is Critical:** Bot behavior evolves rapidly; models must adapt dynamically through retraining and feedback loops.
3. **Operational Integration Matters:** AI must be embedded into real-time pipelines and moderation workflows to maximize impact.
4. **Balance Between Security and User Experience:** False positives can undermine trust. Thresholds, verification prompts, and system transparency must be carefully managed.

4.8. Broader Implications

The successful deployment of AI in SOYO illustrates broader implications for social applications in similar markets. Platforms can achieve significant reductions in bot activity, enhance user experience, and reduce operational costs. Moreover, this approach supports regulatory compliance by proactively addressing fraudulent activity and protecting user data.

AI-driven security systems, as demonstrated by SOYO, also contribute to academic knowledge in social computing, cybersecurity, and applied machine learning. By documenting real-world deployment, performance metrics, and system design choices, this case study provides valuable evidence for the effectiveness of hybrid AI interventions in social ecosystems.

V. Conclusion

The deployment of a hybrid AI system on the SOYO social platform demonstrates the practical effectiveness of combining multiple artificial intelligence techniques to mitigate bot activity, protect user data, and maintain platform integrity. By integrating **NLP-based message content analysis, behavioral anomaly detection, and CNN-based face verification**, SOYO successfully reduced bot-generated messages by **96%** and decreased user complaints from 100% to **13%**, illustrating a significant improvement in both system performance and user experience.

This case study highlights several key insights. First, **multi-layered AI interventions** are essential in dynamic social environments. While single-method solutions may be effective against specific attack vectors, they are often insufficient against adaptive, coordinated bot networks. The synergy between content filtering, behavioral monitoring, and identity verification creates a resilient framework capable of addressing diverse threats.

Second, the importance of **real-time processing and scalability** cannot be overstated. SOYO's architecture, built on streaming pipelines and optimized neural network deployments, enabled instantaneous detection and mitigation of malicious activities. This responsiveness not only enhances security but also reinforces user trust by minimizing exposure to fraudulent interactions.

Third, **continuous learning mechanisms** are critical for maintaining long-term effectiveness. Bot strategies evolve rapidly, and AI models must adapt through retraining, feedback incorporation, and threshold adjustment. SOYO's approach demonstrates that hybrid AI systems can remain effective in the face of evolving threats when appropriately maintained and monitored.

Moreover, the implementation underscores the significance of **balancing security with user experience**. While aggressive bot detection improves platform safety, false positives can negatively impact genuine users. SOYO addressed this by optimizing thresholds, providing clear verification prompts, and integrating human oversight where necessary. Such practices ensure that security interventions enhance rather than detract from user engagement.

Finally, the SOYO experience offers broader implications for social platforms, particularly those operating in emerging markets. The combination of AI-driven detection, operational integration, and ethical data handling provides a

replicable model for other applications seeking to reduce bot activity while maintaining scalability and user trust. This approach contributes not only to platform performance but also to the academic understanding of AI in social computing, cybersecurity, and applied machine learning.

In conclusion, SOYO's hybrid AI system serves as a robust, adaptable, and scalable solution for bot mitigation in social applications. By leveraging advanced AI techniques, the platform has effectively safeguarded users, optimized operational efficiency, and set a benchmark for the deployment of AI-driven security in social networks. Future work may focus on integrating more advanced generative AI detection, further reducing false positives, and exploring cross-platform applications of similar hybrid AI frameworks.

Acknowledgements

The author would like to thank the SOYO development and moderation teams for providing access to operational insights and anonymized user activity data that enabled this study. Special thanks to the AI engineering team for sharing implementation details regarding the NLP classifier, behavioral anomaly detection, and CNN-based face verification modules.

The author also acknowledges the University of Florida, Warrington College of Business, for academic support, guidance in data analysis, and the facilitation of research resources. This work was conducted as part of the author's master's degree in Information Systems and Operations Management, with a focus on data science applications in real-world social platforms.

References

[1]. Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. *ACM Computing Surveys, 50*(1), 1-39.

[2]. Zhang, Y., Yu, H., Chen, Y., & Li, Q. (2020). Bot detection in social networks: A deep learning approach. *IEEE Transactions on Information Forensics and Security, 15*, 2873–2886.

[3]. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

[4]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

[5]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[6]. SOYO Platform Data. (2025). Internal anonymized user activity and moderation reports, Kenya and Nigeria operations (January–March 2025).

[7]. Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

[8]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58.

[9]. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.