AI-Driven Techniques for Detecting Malicious and Fake Profiles Across Social Media Networks

[1]*Mutyala srisailakshmi*
[1]*M.Tech, Department of Computer Science and Engineering*
DNR College of Engineering & Technology (Autonomous)
Bhimavaram – 534202, Andhra Pradesh, India
srisailakshmi068@ gmail. Com

[2]*Dr.B.V.S.Varma*
[2]*Professor, Department of Computer Science and Engineering*
DNR College of Engineering & Technology (Autonomous)
Bhimavaram – 534202, Andhra Pradesh, India
phdvarma@gmail.com

Abstract—The rapid growth of social media platforms has transformed global communication but has also enabled a sharp rise in fake and malicious accounts that spread spam, misinformation, and phishing attacks, thereby undermining user trust and platform integrity. Traditional rule-based and manual moderation techniques are inadequate for identifying such accounts across large-scale, rapidly evolving social networks. This project proposes an AI-driven system for detecting malicious and fake profiles by analyzing multiple behavioral and structural features, including account activity patterns, posting frequency, follower-to-following ratio, and profile metadata. Machine learning algorithms such as Random Forest, Support Vector Machines (SVM), and Neural Networks are employed to classify accounts as genuine or fake, while deep learning models capture complex, nonlinear behavioral patterns that traditional classifiers may overlook. Natural Language Processing (NLP) techniques are further applied to analyze the textual content of posts for spam indicators and malicious intent. Ensemble learning methods combine multiple classifiers to improve overall detection accuracy and robustness. The proposed system supports real-time detection to curb the spread of misinformation, along with automated alerting and reporting of suspicious accounts. Cloud-based deployment ensures scalability to accommodate millions of users, while visualization dashboards provide analytics on detected threats. Adaptive learning mechanisms allow the model to evolve alongside changing fake-account strategies. System performance is evaluated using accuracy, precision, recall, and F1-score. Overall, the proposed AI-driven framework offers a reliable, scalable, and proactive approach to enhancing social media security.

Keywords— Fake Profile Detection, Malicious Account Detection, Machine Learning, Deep Learning, Natural Language Processing, Random Forest, Support Vector Machine, Neural Networks, Ensemble Learning, Social Media Security, Spam Detection, Real-Time Detection, Adaptive Learning.

I. INTRODUCTION

The rapid expansion of social media platforms has transformed global communication, enabling billions of users to connect, share information, and conduct business online. However, this growth has also created opportunities for the creation of fake and malicious accounts that spread spam, misinformation, and phishing attacks, thereby undermining trust in online platforms. Fake profiles are often used to manipulate public opinion, inflate follower counts, conduct financial fraud, or impersonate real individuals and organizations [1]–[3]. As these platforms scale to millions of active users, the accurate and timely identification of such accounts has become a critical security requirement.

Conventional fake-account detection generally relies on manual moderation, user reporting, and static rule-based filtering, such as blacklisting known spam keywords or flagging accounts that exceed fixed activity thresholds. Although these methods provide a baseline level of protection, they have limited capability for continuous, large-scale analysis and cannot adapt to the constantly evolving strategies employed by malicious actors [2], [3]. Machine-learning techniques can learn relationships among behavioral, structural, and textual account features to distinguish genuine users from fake ones with greater accuracy [4], [5]. Deep-learning and hybrid neural-network models can further capture complex, nonlinear behavioral patterns that traditional rule-based or shallow classifiers often fail to detect [6], [7].

Fake-account detection is further complicated because malicious accounts continuously change their behavior to evade detection. Ensemble learning and adaptive-learning approaches have therefore been explored to combine multiple classifiers and update detection models over time [8]–[10]. Natural Language Processing (NLP) techniques have also improved the identification of spam, phishing, and malicious intent embedded in post and bio content [9], [11]. However, class imbalance between the small number of labeled fake accounts and the much larger number of genuine accounts, limited and outdated datasets, evolving evasion strategies, and computational costs at platform scale remain significant challenges [12], [13].

Detection alone cannot provide complete protection for social media platforms. A practical AI-based system should integrate data acquisition, preprocessing, feature extraction, classification, real-time alerting, and reporting into a unified security pipeline. This project therefore aims to:

1. Review machine-learning and deep-learning approaches for detecting fake and malicious social media profiles.
2. Examine behavioral, structural, and metadata-based features used for account classification.
3. Analyze NLP-based techniques for identifying spam and malicious intent in textual content.
4. Compare the strengths, limitations, and applicability of existing AI-based detection methods.
5. Examine ensemble-learning and adaptive-learning strategies for improving detection robustness.
6. Formulate an integrated AI-based fake-profile detection and reporting framework.
7. Identify evaluation requirements, research challenges, and future directions for reliable and scalable social media security systems.

The remainder of this paper discusses existing detection approaches, the proposed comprehensive framework, evaluation requirements, challenges, future research directions, and concluding observations.

II. EXISTING METHODS FOR AI-BASED FAKE AND MALICIOUS PROFILE DETECTION

Existing fake-profile detection approaches can be broadly classified into manual moderation, feature-based monitoring, conventional machine learning, deep learning, hybrid/ensemble detection, and NLP-based content analysis.

A. Manual Moderation and Rule-Based Detection

Manual moderation depends on user reports, administrator review, and static rule-based filters such as keyword blacklists or fixed activity thresholds. It can provide direct human judgment for flagged content, but frequent review becomes difficult at the scale of modern social networks. Delayed responses and high operational effort limit the early identification of large-scale fake-account campaigns [1], [3].

B. Feature-Based Account Monitoring

Feature-based systems continuously collect account-level indicators such as follower/following ratio, posting frequency, account age, and profile completeness. These systems support more frequent, automated observation than manual review and generate historical datasets for predictive analysis. However, incomplete metadata, privacy restrictions, and inconsistent platform APIs can affect data reliability and downstream model performance [2], [5].

C. Conventional Machine Learning

Machine-learning approaches learn relationships between account features and account authenticity from labeled historical data. Algorithms such as Random Forest, Support Vector Machines, and other supervised techniques are widely used for fake-account classification [1], [4], [6]. Their advantages include automated analysis and relatively efficient computation. However, performance depends heavily on feature selection, dataset quality, and class balance between genuine and fake accounts [12].

D. Deep Learning-Based Detection

Deep-learning models can capture complex, nonlinear relationships among multiple behavioral and textual account variables. Recurrent and transformer-based architectures have been investigated for modeling temporal posting behavior and sequential activity patterns [8]–[10]. These approaches can provide improved representation of subtle behavioral cues, but they generally require large labeled datasets and greater computational resources.

E. Hybrid and Ensemble Detection

Hybrid and ensemble models combine multiple classifiers or learning mechanisms to improve overall detection accuracy and robustness. Recent studies have investigated ensemble methods and reinforcement-learning-based approaches for detecting spam bots and coordinated malicious accounts [11], [15], [16]. Such models can reduce false positives and adapt to new fake-account strategies, although their effectiveness can depend on classifier diversity and training-data currency.

F. NLP-Based Content Analysis

Natural Language Processing techniques analyze the textual content of posts, comments, and bio fields to identify spam, phishing links, and malicious intent. Machine learning can transform textual features into vector representations for classification alongside behavioral features [9], [11]. Although such approaches improve detection coverage beyond purely numerical features, practical challenges include multilingual content, sarcasm, evolving slang, and adversarial text manipulation.

TABLE I: COMPARISON OF EXISTING FAKE-PROFILE DETECTION METHODS

Method	Main Technique	Strength	Major Limitation
Manual moderation	Human review/reporting	Direct judgment	Time-consuming, not scalable
Feature-based monitoring	Metadata/activity tracking	Frequent, automated data	Incomplete/API-restricted data
Machine learning	RF, SVM, etc.	Automated classification	Data/class-imbalance dependent
Deep learning	RNN, Transformers	Captures complex behavior patterns	Requires large labeled data
Hybrid/ensemble	Combined classifiers	Improved accuracy, adaptability	Increased model complexity
NLP-based	Text/sentiment analysis	Detects spam/malicious intent	Language/adversarial-text limits
Proposed framework	**Integrated AI (ML+DL+NLP+Ensemble)**	**Detection + real-time alerting + adaptive learning**	**Requires comprehensive validation**

The literature therefore indicates a transition from manual, rule-based moderation toward continuous, automated, and intelligent fake-profile detection. However, many existing approaches focus primarily on a single feature type or detection technique. A comprehensive architecture should integrate behavioral, structural, and textual data acquisition, preprocessing, multi-model classification, and real-time reporting into a unified security pipeline.

III. PROPOSED AI-POWERED FAKE PROFILE DETECTION AND MANAGEMENT SYSTEM

The proposed system considers social media account security as an integrated sequence of interconnected processing stages:

Social Media Accounts + User Activity Data
↓
Data Acquisition
↓
Data Preprocessing
↓
Behavioral, Structural, and Textual Feature Extraction
↓
AI-Based Classification (ML/DL/NLP + Ensemble)
↓
Fake/Malicious Account Detection
↓
Risk Scoring and Anomaly Assessment
↓
Automated Alerting and Reporting
↓
Platform Administrator Review and Action

The architecture consists of five major functional layers: data acquisition, preprocessing, intelligent analysis, classification and risk assessment, and decision support. The first layer collects important account-level parameters such as follower/following ratio, post frequency, account age, profile metadata, and textual content. The preprocessing layer removes noise, handles missing values, and converts heterogeneous account data into suitable input features.

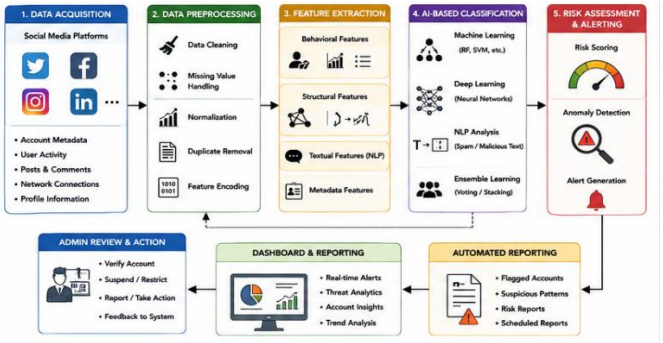

Fig. 1. Overall Architecture of the Proposed AI-Powered Fake Profile Detection System

The intelligent analysis layer applies machine-learning, deep-learning, and NLP techniques to identify abnormal behavioral patterns and suspicious textual content. The classification layer estimates the likelihood that an account is genuine or fake based on historical and real-time activity. The risk-assessment layer combines classification outputs to generate actionable alerts for potentially malicious accounts.

Finally, the decision-support layer presents predictions, detected anomalies, and recommended actions to platform administrators through an accessible monitoring dashboard. The architecture is designed as a decision-support framework, assisting moderators in timely intervention rather than completely replacing human judgement. This integrated approach enables continuous monitoring, early identification of threats, adaptive detection, and more efficient social media security operations.

IV. FAKE PROFILE DATA AND ACCOUNT REPRESENTATION

A. Account Data Acquisition and Processing

The proposed system collects account parameters such as follower count, following count, post frequency, account age, profile completeness, and textual content from posts and bios. Preprocessing handles missing, duplicate, inconsistent, and abnormal data, and normalizes numerical values.

The structured account representation is defined as:

$$A_i = \{B_i, S_i, T_i, M_i\} \quad (2)$$

where B_i represents behavioral features, S_i represents structural/network features, T_i represents textual content features, and M_i represents profile metadata.

Behavioral, structural, textual, and metadata indicators are analyzed to assess account authenticity. Normalization converts parameters with different scales and units into comparable ranges, improving joint model learning and classification.

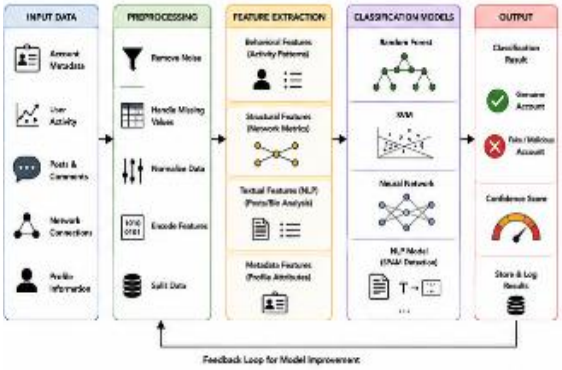

Fig. 2. Account Data Processing and Fake Profile Classification Pipeline

B. Account Condition Representation

The behavioral and structural condition of a social media account is represented using relevant activity and profile parameters. The system extracts:

- Follower/following ratio
- Post frequency
- Account age
- Profile completeness (bio, photo, links)
- Posting-time regularity
- Language/content patterns
- Duplicate or spam content indicators
- Network connections and interaction patterns

The account condition representation is defined as:

$$Q = \{FR, PF, AA, PC, PR, L, D, N\} \quad (2)$$

Where FR represents follower/following ratio, PF represents post frequency, AAAA AA represents account age, PC represents profile completeness, PR represents posting-time regularity, L represents language/content patterns, D represents duplicate/spam indicators, and N represents network interaction features.

Representing the collected account parameters using a unified structure enables the proposed system to analyze the relationship between behavioral activity, textual content, and network structure. The resulting representation can subsequently be provided to the machine-learning or deep-learning model for account authenticity prediction, anomaly detection, or malicious-intent assessment, depending on the objective of the proposed system.

V. SEMANTIC BEHAVIOR–POLICY MATCHING AND ACCOUNT RISK SCORING

A central component of the proposed system is semantic behavior–policy matching. Unlike simple threshold-based analysis, the proposed approach identifies relationships between an account's observed behavior and known patterns associated with genuine versus malicious activity. The system considers parameters such as posting frequency, follower/following ratio, bio and content language, account age, and other available profile parameters to determine the likelihood that an account is fake or malicious.

The observed account behavior and the reference patterns of known genuine/fake accounts are each converted into feature embeddings, and their semantic similarity is computed to measure how closely an account's behavior aligns with established genuine-user or fake-account patterns. A higher similarity to fake-account patterns indicates a greater likelihood that the account is malicious, while a higher similarity to genuine-user patterns indicates a lower risk.

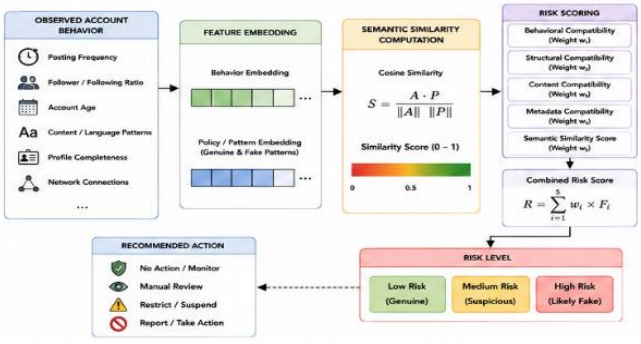

Fig. 3. Semantic Behavior–Policy Matching and Account Risk Scoring

Semantic similarity alone, however, should not determine the final classification decision. An account with an overall behavior pattern similar to genuine users may still contain a critical red flag, such as a sudden spike in posting frequency or a suspicious follower/following ratio, that can indicate malicious intent. Therefore, the proposed system combines semantic similarity with explicit, account-specific risk factors, including:

- Behavioral compatibility (activity pattern consistency)
- Structural compatibility (follower/following ratio, network density)
- Content compatibility (spam or malicious-intent indicators in posts/bio)
- Metadata compatibility (profile completeness, account age)
- Semantic similarity to known genuine/fake behavior patterns

This weighted combination allows the fake-profile detection system to assign different levels of importance to different indicators depending on the platform and threat type. For example, a spam-detection application may emphasize content and posting-frequency indicators, whereas a bot-detection application may emphasize structural and network-based indicators.

The accounts are subsequently evaluated according to their combined risk assessment to generate an account authenticity status and corresponding recommended action, enabling timely intervention and improved social media security management.

VI. EXPLAINABLE AI AND DECISION-AWARE FAKE PROFILE MANAGEMENT

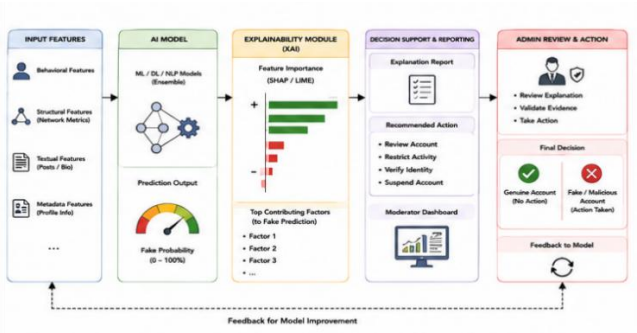

Fig. 4. Explainable AI Framework for Fake Profile Detection and Decision Support

AI-based fake-profile detection systems should provide more than a binary prediction or numerical score. Platform administrators and moderators need to understand why the system flagged a particular account as suspicious or fake. Explainable AI (XAI) techniques such as SHAP and LIME can provide feature-level explanations of model predictions.

For each prediction, the proposed system can generate an explanation containing:

1. Behavioral-pattern contribution
2. Follower/following ratio influence
3. Posting-frequency influence
4. Textual-content (spam/malicious-intent) contribution
5. Profile metadata and completeness influence
6. Network/connection-pattern indicators
7. Major factors contributing to the final prediction
8. Recommended moderation action

For example:

Account Risk Assessment — 91% Likely Fake

- Behavioral pattern: highly irregular posting activity.
- Follower/following ratio: abnormally skewed.
- Content: contains repeated spam-like links.
- Profile metadata: incomplete bio, no verified information.
- Network pattern: connected to other flagged accounts.
- Major influencing factor: coordinated spam-content pattern.
- Recommendation: Flag account for review and restrict posting privileges pending verification.

Such explanations help administrators understand whether the AI prediction is supported by meaningful account indicators rather than relying only on the final classification score.

A. Explainability and Decision Monitoring

Fake-profile classification models can be influenced by multiple behavioral, structural, and textual parameters. Therefore, the system should continuously monitor the contribution of each input feature to the prediction. Explainability helps identify which parameters have the greatest influence on flagging an account as fake, spam-related, or malicious.

The proposed framework therefore integrates explainability as a decision-support component. The system should report both:

Performance = How accurately does the system detect fake or malicious accounts?

Explainability = Can the administrator understand which account behaviors caused the flag?

This dual objective is important because a highly accurate model that provides no understandable reasoning may be difficult to trust or act upon in practical content-moderation environments.

VII. ADAPTIVE INTELLIGENCE AND HUMAN-IN-THE-LOOP DECISION SUPPORT

Fake-account strategies are not static. Spam techniques, bot behaviors, follower/following patterns, and content-evasion tactics continuously evolve as malicious actors adapt to existing detection systems. A fake-profile detection model trained using fixed behavioral patterns may therefore become less effective as new evasion strategies emerge.

Model Update

Adaptive mechanisms can update behavioral thresholds, feature importance, detection knowledge, and moderation recommendations while maintaining model version control.

The system can periodically incorporate newly collected account data to improve its understanding of evolving fake-account strategies. However, adaptive AI should not automatically modify critical moderation actions, such as account suspension, without validation. Model updates should first be evaluated using historical and newly collected labeled datasets before being deployed.

Human-in-the-Loop Workflow

The final fake-profile decision-support workflow is:

Data Collection → AI Prediction → Risk Assessment → Explanation → Administrator Verification → Moderation Recommendation → Final Administrator Decision

The platform administrator or moderation team remains responsible for the final decision, particularly when account behavior is ambiguous or when additional contextual information is required that may not be available in the collected data.

This human-in-the-loop design allows administrators to combine AI-based predictions with platform policy knowledge

and real-time contextual observations. Administrator feedback can also be captured and used for future system improvement, while preventing the AI system from making completely autonomous account-suspension decisions without human supervision.

VIII. EVALUATION FRAMEWORK

A comprehensive evaluation framework is adopted to assess the proposed fake-profile detection system across classification performance, semantic matching, explainability, fairness, efficiency, and robustness. Rather than relying on a single performance measure, the evaluation considers multiple complementary dimensions to determine the practical effectiveness of the proposed framework.

A. Classification Performance

The classification module is evaluated using accuracy, precision, recall, and F1-score. Accuracy measures overall classification correctness, while precision indicates the proportion of accounts flagged as fake that are actually fake. Recall measures the ability to correctly identify genuinely fake accounts, and F1-score provides a balanced assessment of precision and recall. These metrics are important because both false accusations of genuine users and missed fake accounts can undermine platform trust and security.

B. Detection Ranking Performance

Since prioritizing the most suspicious accounts for review is important at scale, the system is evaluated using Precision@K, Recall@K, and risk-score ranking consistency. These metrics determine whether the most likely fake or malicious accounts are positioned at the top of the moderation queue, ensuring limited moderator resources are used effectively.

C. Semantic Matching

The semantic behavior-matching component is compared with a conventional threshold-based baseline. The evaluation examines whether contextual behavioral representations can identify malicious accounts despite variations in evasion tactics, platform type, and account presentation. This assessment determines the effectiveness of contextual account-risk alignment.

D. Explainability

The generated account-flagging explanations are evaluated based on faithfulness, consistency, completeness, and human understandability. The explanation should accurately represent the factors influencing the classification decision and enable moderators to verify behavioral, structural, and textual evidence supporting the flag.

E. Fairness

Fairness evaluation examines potential differences in flagging outcomes across legitimate user groups, such as new accounts, accounts from particular regions, or accounts with limited activity history. Controlled testing can determine whether the system disproportionately flags genuine users from certain groups, ensuring that detection primarily reflects malicious behavior rather than unrelated account characteristics.

F. Efficiency and Robustness

Computational evaluation considers account-processing time, classification latency, memory consumption, throughput, and scalability to millions of accounts. Robustness is evaluated through cross-platform testing, evolving spam/bot strategies, incomplete profile information, and diverse account types. This ensures that the proposed system maintains reliable performance under practical, large-scale social media conditions.

TABLE II: EVALUATION DIMENSIONS OF THE PROPOSED FAKE PROFILE DETECTION FRAMEWORK

Dimension	Evaluation Metrics / Criteria	Purpose
Classification	Accuracy, Precision, Recall, F1-Score	Evaluate fake-account identification performance
Detection Ranking	Precision@K, Recall@K, Risk-score ranking	Prioritize highest-risk accounts for review
Semantic Matching	Semantic similarity, Threshold-based baseline comparison	Assess behavioral risk alignment
Explainability	Faithfulness, Consistency, Completeness, Human Understandability	Evaluate decision transparency
Fairness	Group disparity, Controlled account-group evaluation	Identify potential moderation bias
Efficiency	Processing Time, Classification Latency, Memory, Throughput	Assess computational efficiency at scale
Robustness	Cross-platform testing, Evasion-strategy variation, Incomplete-data testing	Evaluate generalization capability

IX. RESEARCH CHALLENGES AND FUTURE ROADMAP

Although AI-powered detection can improve fake and malicious profile identification, several challenges remain in developing a reliable and practical system.

A. Dataset Availability

Fake-profile datasets may contain limited labeled examples, outdated fake-account patterns, and variations across platforms. Developing sufficiently diverse and representative, up-to-date datasets is therefore important for reliable detection.

B. Evolving Evasion Strategies

Malicious actors continuously adapt their behavior, content, and network patterns to evade detection. The proposed system must therefore handle variations in spam tactics, bot behaviors, and coordinated inauthentic activity to achieve reliable, long-term detection.

C. Detection Reliability

Account classification should consider multiple behavioral, structural, and textual factors rather than depending on a single indicator. Future improvements should focus on maintaining consistent detection accuracy across different platforms and account types.

D. Explainability

AI-generated fake-account flags should provide understandable reasons for moderation decisions. Future systems should improve the transparency of behavioral-pattern matching, content-risk relevance, and network-based evidence.

E. Fairness

Detection systems should minimize inappropriate influence from unrelated account attributes, such as region, language, or new-user status. Fairness-aware evaluation is therefore required to ensure that flagging primarily reflects malicious behavior rather than incidental account characteristics.

F. Platform and Threat Evolution

New platforms, features, and malicious techniques continuously emerge. Future detection systems should therefore support updated behavioral representations and evolving threat patterns without reducing previously learned detection capabilities.

G. Human–AI Interaction

AI recommendations should assist moderators rather than completely replace human judgement. Future systems should provide effective moderator feedback mechanisms and

decision-support interfaces for reviewing and validating flagged accounts.

H. Privacy and Security

Account and user data contain personal and behavioral details. Secure data management, controlled access, anonymization, and privacy-preserving processing are therefore important for practical, ethical deployment of fake-profile detection systems. The research roadmap can therefore be organized into three stages:

Short Term: Improve feature extraction, semantic behavior matching, multi-factor classification, explainability, and fairness evaluation.

Medium Term: Develop adaptive behavioral representations, improved platform-specific detection, diverse and updated datasets, and robust cross-platform evaluation.

Long Term: Develop reliable, human-centered social media security intelligence that combines semantic matching, explainable classification, fairness-aware evaluation, adaptive learning, and human-in-the-loop moderation support.

X. CONCLUSION

This paper presented an AI-powered framework for intelligent detection of fake and malicious social media profiles. The framework addresses the limitations of conventional manual moderation and rule-based detection systems by incorporating behavioral, structural, and textual data processing, semantic behavior-risk matching, multi-factor account scoring, real-time detection, explainability, fairness monitoring, and human-in-the-loop decision support.

The proposed approach evaluates accounts using activity patterns, follower/following ratios, profile metadata, and textual content, combining Random Forest, Support Vector Machines, Neural Networks, and NLP-based techniques with ensemble learning to improve classification accuracy. Semantic matching enables the system to identify contextual relationships between observed account behavior and known genuine or malicious patterns beyond simple threshold-based rules. The risk-scoring component further prioritizes accounts according to their overall likelihood of being fake or malicious.

Explainability and fairness are incorporated to improve the transparency and reliability of detection outcomes. Administrators can examine the major factors contributing to account flags, while fairness evaluation helps identify potentially inappropriate influences on moderation outcomes. The human-in-the-loop design ensures that AI-generated recommendations support moderator decision-making rather than completely replacing human judgement.

Future research should focus on diverse and updated datasets, adaptive behavioral representations, robust cross-platform evaluation, faithful explainability, fairness-aware learning, privacy-preserving processing, and reliable AI-based social media security solutions.

REFERENCES

[1] N. Singh, T. Sharma, A. Thakral, and T. Choudhury, "Detection of Fake Profile in Online Social Networks Using Machine Learning," in *Proc. 2018 Int. Conf. Advances in Computing and Communication Engineering (ICACCE)*, Paris, France, 2018, pp. 231–234, doi: 10.1109/ICACCE.2018.8441713.

[2] S. Khaled, N. El-Tazi, and H. M. O. Mokhtar, "Detecting Fake Accounts on Social Media," in *Proc. 2018 IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 3672–3681, doi: 10.1109/BigData.2018.8621913.

[3] M. Conti, R. Poovendran, and M. Secchiero, "FakeBook: Detecting Fake Profiles in On-Line Social Networks," in *Proc. 2012 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM)*, 2012, pp. 1071–1078, doi: 10.1109/ASONAM.2012.185.

[4] E. Van Der Walt and J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018, doi: 10.1109/ACCESS.2018.2796018.

[5] J. Jia, B. Wang, and N. Z. Gong, "Random Walk Based Fake Account Detection in Online Social Networks," in *Proc. 47th Annual IEEE/IFIP Int. Conf. Dependable Systems and Networks (DSN)*, 2017, pp. 273–284, doi: 10.1109/DSN.2017.60.

[6] E. Anupriya et al., "Fraud Account Detection on Social Network Using Machine Learning Techniques," in *Proc. IEEE Conf.*, 2022, doi: 10.1109/ICMNWC56175.2022.10088336.

[7] N. Kumar and D. Poonam, "Detection and Prevention of Profile Cloning in Online Social Networks," in *Proc. 2019 5th Int. Conf. Signal Processing, Computing and Control (ISPCC)*, 2019, pp. 287–291, doi: 10.1109/ISPCC48220.2019.8988385.

[8] D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. B. Hernández, A. Lozano-Diez, and F. Álvarez, "A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers," *IEEE Access*, vol. 9, pp. 54591–54601, 2021, doi: 10.1109/ACCESS.2021.3068659.

[9] M. Fazil, A. K. Sah, and M. Abulaish, "DeepSBD: A Deep Neural Network Model with Attention Mechanism for SocialBot Detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4211–4223, 2021, doi: 10.1109/TIFS.2021.3097146.

[10] E. Arin and M. Kutlu, "Deep Learning Based Social Bot Detection on Twitter," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1763–1772, 2023, doi: 10.1109/TIFS.2023.3254429.

[11] H. Ping and S. Qin, "A Social Bots Detection Model Based on Deep Learning Algorithm," in *Proc. 2018 IEEE 18th Int. Conf. Communication Technology (ICCT)*, 2018, pp. 1435–1439.

[12] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[13] S. R. Sahoo and B. B. Gupta, "Classification of Various Attacks and Their Defence Mechanism in Online Social Networks: A Survey," *IEEE Access*, related survey, 2019.

[14] M. Al-Qurishi, M. AlRubaian, S. M. M. Rahman, A. Alamri, and M. M. Hassan, "A Prediction System of Sybil Attack in Social Network Using Deep-Regression Model," *Future Generation Computer Systems*, IEEE-affiliated study, 2018.

[15] G. Lingam, R. R. Rout, and D. V. L. N. Somayajulu, "Particle Swarm Optimization on Deep Reinforcement Learning for Detecting Social Spam Bots and Spam-Influential Users in Twitter Network," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2281–2292, 2021, doi: 10.1109/JSYST.2020.2986268.