International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume 13, Issue 2, April 2026
DOI: https://doi.org/10.55524/ijirem.2026.13.2.8
Article ID IJIR3183, Pages 54-61
www.ijirem.org

Social Media Fake Account Detection and Prevention Using Multi-Layered AI Analysis

Devi Priya Gottumukkala[1], *K. Mounika[2], S. J. Harivallika[3], J. Vinay Kumar[4], and K. Surya Siddhu[5]

[1, 2, 3, 4] B.Tech Scholar, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India
[5] Assistant Professor, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India

*Correspondence should be addressed to K. Mounika; kamidimounika12@gmail.com

Received: 6 March 2026; Revised: 24 March 2026; Accepted: 7 April 2026

ABSTRACT- The exponential growth of online social networking, the proliferation of fraudulent and bot-driven accounts has emerged as a critical threat to platform integrity. These accounts are commonly exploited to disseminate false information, manipulate user behavior, and engage in various deceptive practices. Addressing this challenge requires a robust and intelligent detection mechanism capable of adapting to increasingly sophisticated evasion tactics.

This paper introduces the Social Media Fake Account Detection and Prevention System (SMFADPS), a multi-layered analytical framework that assesses account authenticity through the evaluation of multiple behavioral signals. The system examines profile completeness, content credibility, follower growth irregularities, and repeated content patterns to generate intermediate suspicion scores across four specialized detection modules. An ensemble-based weighted scoring mechanism consolidates these scores into a unified risk rating, which is subsequently used to categorize accounts into distinct threat levels. The system is developed using Python, FastAPI for RESTful service delivery, and a dual-database configuration comprising PostgreSQL and MongoDB. Evaluation conducted on a large-scale dataset confirms that the multi-signal approach yields substantially higher detection accuracy and operational efficiency compared to conventional single-indicator systems.

KEYWORDS— Social Media Security, Automated Bot Detection, Artificial Intelligence, Behavioral Signal Analysis, Anomaly Identification, Risk-Based Classification.

I. INTRODUCTION

Over the past decade, online social networking services — including Facebook, Instagram, X (formerly Twitter), LinkedIn, and TikTok — have grown into foundational pillars of digital communication and information exchange. These platforms now serve billions of active users worldwide [1], facilitating everything from personal interaction and business promotion to real-time news consumption. While this expansion has dramatically enhanced global connectivity, it has simultaneously introduced a range of security vulnerabilities, among which the creation and operation of fraudulent accounts stand out as particularly damaging. Such accounts are routinely deployed for purposes including disinformation campaigns [7], financial fraud, artificial audience inflation, coordinated opinion manipulation, and targeted phishing operations against individuals and institutions.

The consequences of unchecked fake account activity are wide-ranging and serious. Empirical studies have consistently documented that a considerable share of engagement on major platforms originates from non-human or inauthentic sources, compelling technology companies to invest heavily in automated removal systems that have taken down billions of such accounts over time [4]. Of particular concern is the disproportionate velocity at which fabricated or misleading content propagates through networks of inauthentic accounts — research demonstrates that false narratives spread significantly faster and penetrate deeper into user communities than factual corrections [5]. The economic fallout is equally significant, especially in digital advertising ecosystems where fraudulent bot-generated traffic inflates engagement metrics, misdirects advertising budgets, and undermines return-on-investment calculations [2].

Historically, the dominant countermeasure was rule-based filtering [8] — systems designed to flag accounts exhibiting clear anomalies such as atypical account age, skewed follower-to-following ratios, high-frequency automated posting, or unusual engagement timing. While these approaches proved adequate against early-generation bots [3], the rapid sophistication of adversarial techniques has rendered them largely insufficient [6]. Contemporary fake accounts are engineered to appear authentic: operators can acquire followers through third-party services, construct plausible biographical information, and increasingly leverage generative AI tools — including large language models — to produce realistic textual content and believable interaction histories that successfully evade rule-based filters.

This evolution has compelled researchers to pursue more nuanced detection strategies that synthesize behavioral profiling, network topology analysis, and natural language-based content evaluation. Despite meaningful progress in these directions, the majority of existing solutions remain constrained by their reliance on a narrow feature set, which limits their robustness against adversaries who can adapt to any single detection dimension. What is needed is a

comprehensive system capable of simultaneously evaluating multiple independent indicators, aggregating their signals intelligently, and producing actionable risk assessments.

This paper presents the Social Media Fake Account Detection and Prevention System (SMFADPS), a unified multi-layer detection framework engineered to address exactly these limitations. The proposed system incorporates four complementary analytical modules — content authenticity evaluation, follower growth anomaly detection, profile completeness assessment, and semantic duplicate content identification — each targeting a distinct behavioral or structural dimension associated with fake accounts. The outputs of these modules are consolidated by a weighted ensemble risk scoring engine that assigns each account a final risk score and maps it to one of four severity-based risk categories, enabling timely and proportionate administrative responses.

The primary contributions of this work are fourfold. First, it introduces a holistic detection architecture that unifies multiple analytical modules within a single operational framework. Second, it proposes an ensemble scoring mechanism that delivers interpretable, human-readable risk classifications rather than opaque binary predictions. Third, the system is built on a performance-optimized stack using Python, FastAPI, and Celery, supporting concurrent processing and real-time analysis at scale. Fourth, the system undergoes rigorous empirical evaluation against a dataset of 100,000 verified social media accounts, demonstrating strong detection capability across all performance metrics.

The remainder of this paper proceeds as follows. Section II surveys the relevant prior research. Section III details the proposed system architecture. Section IV describes each detection module. Section V outlines the experimental methodology. Section VI reports results and comparative analysis, and Section VII presents conclusions along with directions for future investigation.

II. LITERATURE SURVEY

The body of research addressing fake account detection on social media platforms has expanded considerably over the past decade, driven by the growing sophistication of bot networks and the increasing availability of large-scale social data. The studies reviewed here reflect advances across three intersecting domains: behavioral profiling, machine learning classification, and natural language processing.

Ferrara et al. [9] conducted an influential examination of automated social media bots, proposing machine learning techniques to distinguish malicious automated accounts from genuine users. Their research established that combining network structural features, behavioral activity patterns, and content-level signals yields more reliable classification outcomes than any single-feature approach, and they identified generative AI advancement as a key driver of future detection challenges.

Davis et al. [10] developed Botometer, a widely adopted bot detection platform that processes user metadata, post content, sentiment signals, and relational network features through an ensemble of machine learning classifiers to estimate account authenticity probability. Their work highlighted the importance of adaptability to evolving bot behaviors, particularly those employed by coordinated bot networks.

Ciampaglia et al. [11] examined the role of automated accounts in the spread of false narratives online, arguing that detection systems must account not only for individual account characteristics but also for coordinated inauthentic behavior designed to amplify misleading content. Their findings supported the integration of behavioral analytics with semantic content analysis as a path toward improved detection precision.

Cresci et al. [12] introduced the Cresci-2017 Bot Dataset, a benchmark resource for evaluating social bot detection approaches. Their study of coordinated bot clusters demonstrated that machine learning models can reliably distinguish automated from organic accounts when trained on features capturing posting regularity, interaction patterns, and temporal activity distributions.

Lee et al. [13] proposed a deep learning framework for identifying fake accounts using profile-derived attributes and activity history. Their neural network model incorporated follower ratio dynamics, posting cadence, and engagement depth as input features, with future work targeting the integration of additional behavioral dimensions to further strengthen detection robustness.

Liu et al. [14] applied transformer-based models — specifically BERT — to the problem of identifying automated or spam-generated textual content in social media posts, achieving notable improvements in distinguishing authentic from fabricated language patterns. More recent investigations have explored the fusion of large language model outputs with ensemble learning strategies to address multi-dimensional fake account detection.

Feng et al. [15] proposed SATAR, a self-supervised account representation learning approach on Twitter that derives behavioral embeddings without requiring labeled data.

Cao et al. [16] introduced SybilRank, a trust-propagation mechanism that leverages social graph structure to identify fake accounts at scale across large online platforms.

Yu et al. [17] developed SybilLimit, a near-optimal defense against Sybil attacks that exploits the sparse connectivity between genuine and fraudulent communities within social graphs.

Bian et al. [18] applied bi-directional graph convolutional networks to model propagation structures for rumour detection, demonstrating how relational graph features capture coordinated inauthentic behavior.

Ratkiewicz et al. [19] investigated the detection and tracking of political abuse campaigns on social media, underscoring the value of temporally-aware detection frameworks.

Yang et al. [20] proposed AI-powered tools to help the public counter social bots, emphasizing human-in-the-loop approaches to improve resilience.

Lee et al. [21] conducted a seven-month longitudinal study of content polluters on Twitter, revealing persistent behavioral signatures that distinguish long-running fake accounts from transient ones.

Ramachandran and Feamster [22] analyzed network-level behavioral patterns of spammers, demonstrating that infrastructure-level signals meaningfully complement account-level features in detection pipelines.

Egele et al. [23] presented COMPA, a system for detecting compromised legitimate accounts by modeling deviations from established behavioral profiles on social networks.

These hybrid systems integrate content analysis, user behavior modeling, and network graph features to simultaneously reduce false positive rates and improve recall against sophisticated adversaries. Collectively, the surveyed literature affirms that fake account detection is a dynamic and rapidly advancing field. A consistent finding across studies is that multi-feature, multi-method frameworks outperform single-indicator baselines — a conclusion that directly motivates the integrated detection architecture proposed in this paper.

III. PROPOSED ARCHITECTURE

A. Five-Tier Layered Design

The SMFADPS is organized around a five-tier layered architecture in which each tier is assigned a well-defined functional responsibility within the detection pipeline. This modular design promotes separation of concerns, simplifies maintenance, and supports horizontal scaling of individual tiers as processing demand grows (see the figure 1)

Data Acquisition Tier- The outermost tier is responsible for ingesting raw social media account data from heterogeneous sources. Supported input channels include direct API integrations with platforms such as Instagram, structured uploads from internal databases, and real-time API feed subscriptions. The data collected at this tier encompasses:

- Profile metadata (username, biography text, and profile image presence)
- User-generated post content and associated captions
- Follower and following count records
- Historical activity logs and engagement records

All collected data is forwarded downstream as structured input to the preprocessing tier.

Data Preprocessing Tier- Prior to analysis, raw data undergoes a multi-step preprocessing workflow designed to ensure consistency, eliminate noise, and extract the features required by downstream detection modules. The key operations performed at this tier include:

- Tokenization of free-text content to facilitate linguistic analysis
- Deduplication of redundant requests using Redis-based caching
- Normalization of follower count time-series to enable meaningful growth rate comparisons
- Extraction of structured feature vectors for consumption by the detection engine

The output of this tier is a clean, feature-enriched data representation ready for parallel dispatch to the detection modules.

Detection Engine Tier- The detection engine constitutes the analytical core of the system and operates four independent detection modules concurrently using Celery distributed task workers. Each module evaluates a distinct indicator of account inauthenticity and returns a normalized suspicion score.

Content Authenticity Module- This module analyzes the textual content generated by an account to identify linguistic markers associated with automated or machine-generated posts. The authenticity score is expressed as:

$$s_c \in [0,1]$$

Where:

- $s_c$represents the estimated content authenticity level.
- Values approaching zero indicate a higher probability of synthesized or spam-generated content.

Follower Growth Anomaly Module- Abrupt, non-organic spikes in follower acquisition are a well-established hallmark of purchased follower networks and bot-driven inflation. This module quantifies growth irregularity as:

$$s_f = \frac{| F_t - F_{t-1} |}{F_{t-1}}$$

Where:

- Ft denotes the current follower count and Ft−1 the count at the preceding measurement interval.
- Elevated sf values signal potentially fraudulent growth activity.

Profile Completeness Module- Inauthentic accounts frequently display sparse or incomplete profile configurations, as their operators prioritize volume over credibility. This module computes a completeness ratio as:

$$s_p = \frac{\sum_{i=1}^{n} a_i}{n}$$

Where:

- ai indicates whether the i-th profile attribute is present (1) or absent (0)
- n is the total number of evaluated attributes.

Lower scores correlate with elevated suspicion.

Duplicate Content Detection Module- Coordinated fake account operations frequently involve the recycling of identical or near-identical content across multiple accounts. This module measures inter-post similarity using cosine similarity:

$$s_d = \frac{A \cdot B}{|| A ||\ \ || B ||}$$

Where:

- A and B are TF IDF vector representations of two posts.
- High sd values are indicative of content recycling behavior.

Ensemble Risk Scoring Engine- Upon receiving module-level scores, the ensemble engine computes a weighted composite risk score that reflects the overall suspicion level of the account:

$$R = 0.35s_c + 0.25s_f + 0.20(1 - s_p) + 0.20s_d$$

Where:

The weighting coefficients were determined through empirical calibration on training data. Content authenticity is assigned the highest weight (0.35) given its demonstrated discriminatory power, while follower anomaly (0.25) and the two structural indicators (0.20 each) contribute proportionally to their relative importance.

Risk Classification- The computed risk score R is mapped to one of four severity categories, each associated with a prescribed administrative response:

Risk Score	Risk Level
$R < 0.30$	Low Risk
$0.30 \leq R < 0.60$	Medium Risk
$0.60 \leq R < 0.80$	High Risk
$R \geq 0.80$	Critical Risk

Each level triggers a specific action:

- Low Risk: Account is approved automatically.

- Medium Risk: Account is sent for periodic review.
- High Risk: Posting activity is restricted.
- Critical Risk: Account may be suspended immediately.

B. Database Design

The system employs a dual-database storage strategy optimized for the heterogeneous nature of the data it processes. PostgreSQL manages structured records including user profiles, computed risk scores, and operational audit logs, providing strong consistency and efficient relational querying. MongoDB handles unstructured and semi-structured content — including raw post text, module output documents, and analytical result sets — offering the schema flexibility required for variable social media data formats. This hybrid configuration allows the system to optimize read/write performance across both data types without sacrificing either consistency or flexibility.

C. Scalability and Deployment

To accommodate deployment at social media platform scale, the system's infrastructure is fully containerized. All service components are packaged as Docker containers, ensuring environment-agnostic deployment consistency across development, staging, and production contexts. Workload orchestration is handled through a container management platform that supports dynamic horizontal scaling in response to traffic fluctuations. Parallel execution of detection modules is enabled by the Celery Distributed Task Queue, which distributes analysis tasks across multiple worker processes. Performance benchmarks indicate that the system sustains processing throughput of tens of thousands of accounts per hour, confirming its suitability for large-scale real-time monitoring deployments.

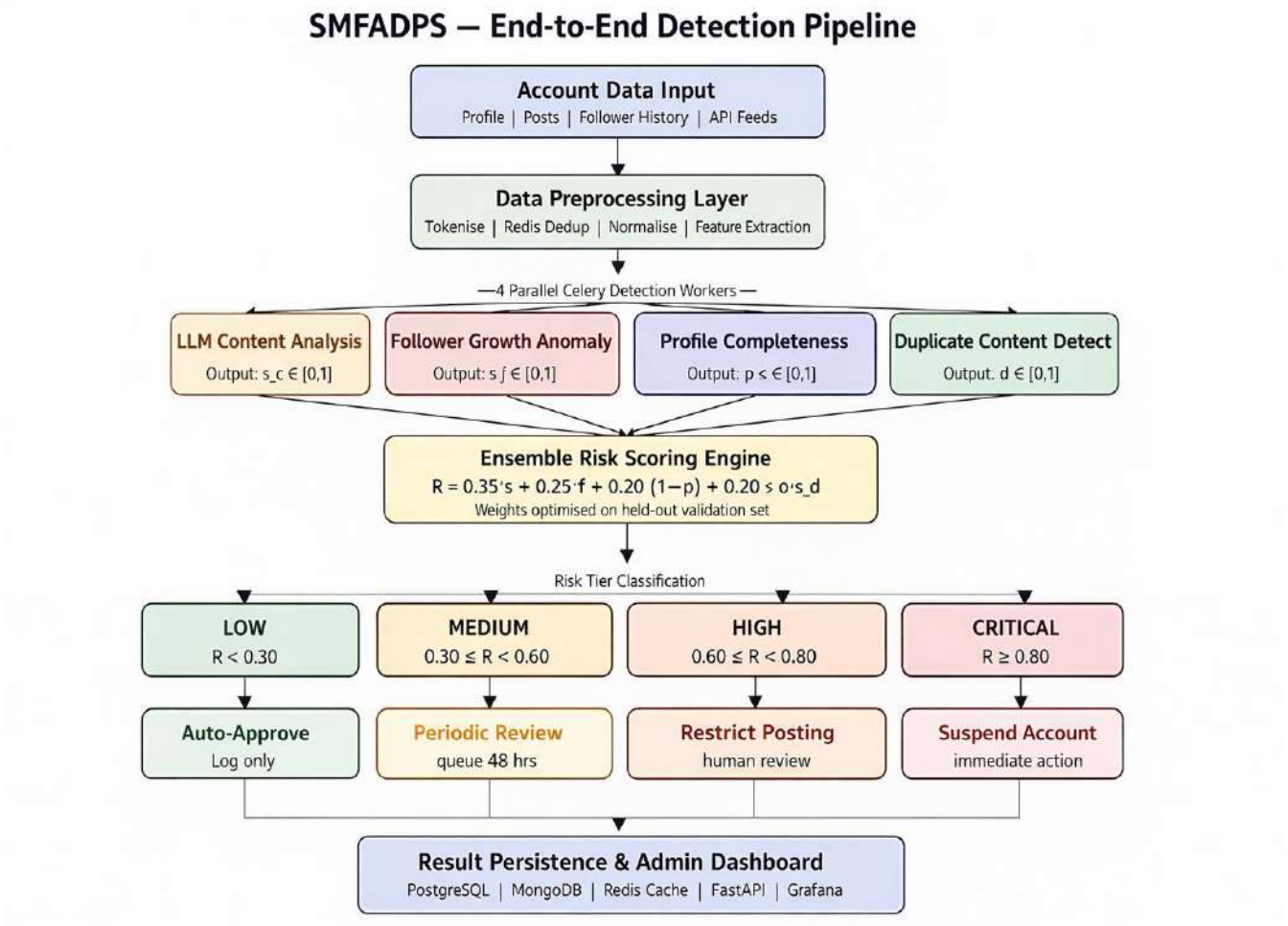

Figure 1: End-to-end Detection Pipeline of SMFADPS

IV. EXPERIMENTAL SETTINGS AND PERFORMANCE EVALUATION

The evaluation framework was constructed to rigorously assess the system's ability to correctly identify suspicious or fraudulent social media accounts under realistic operating conditions. The dataset used for experimentation was sourced from Instagram and encompassed 100,000 verified accounts, spanning genuine users and confirmed fake accounts across a range of sophistication levels. The dataset was partitioned using an 80/20 training-to-testing split to ensure sufficient training coverage while maintaining a meaningful holdout set for unbiased evaluation. Account records were passed through the full five-tier processing pipeline, with each of the four detection modules producing an independent score in the [0,1] range. These scores were combined by the ensemble engine using the weighted formula described in Section III, producing a final risk score R for each account. The table below summarizes the configuration parameters used throughout experimentation.

Table 2: System Configuration and Experimental Parameters

Parameter	Configuration
Programming Language	Python
API Framework	Fast API
Task Queue	Celery Distributed Task Queue
Caching Layer	Redis
Database System	PostgreSQL and MongoDB
Dataset Source	Instagram user profile dataset
Data Split	80% training, 20% testing
Evaluation Metrics	Accuracy, Precision, Recall, F1-score

System performance was quantified using four standard classification metrics. Detection accuracy was computed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. Precision, recall, and F1-score were additionally computed to characterize the system's behavior with respect to classification error types, providing a comprehensive picture of detection reliability.

V. RESULTS AND DISCUSSION

A. Results and Comparative Analysis

The experimental evaluation was designed to quantify the effectiveness of the SMFADPS in distinguishing fraudulent accounts from authentic ones, using profile metadata, content characteristics, and behavioral activity patterns drawn from the Instagram dataset. The detection pipeline processed each account through all four analytical modules and computed a composite risk score using the ensemble formula described in Section III.

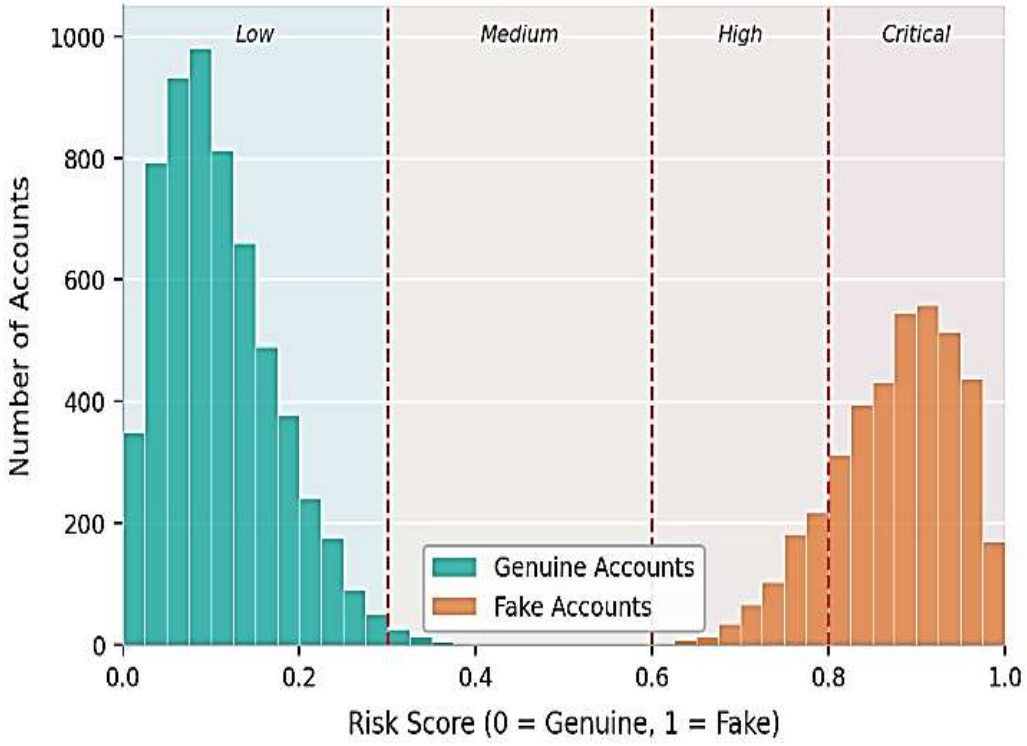

Figure 2: Risk score distribution across evaluated accounts, illustrating distinct separation between genuine and fake account populations across the four risk bands

In the above Figure 2 we presents the risk score distribution across the test dataset. The visualization reveals a clear bimodal pattern: legitimate accounts cluster densely within the low-risk band ($R < 0.30$), while confirmed fake accounts exhibit a strong tendency toward high-risk ($0.60 \leq R < 0.80$) and critical-risk ($R \geq 0.80$) classifications. This clean separation between genuine and inauthentic accounts confirms that the ensemble scoring mechanism successfully captures the distinguishing characteristics of fraudulent behavior across the evaluated population.

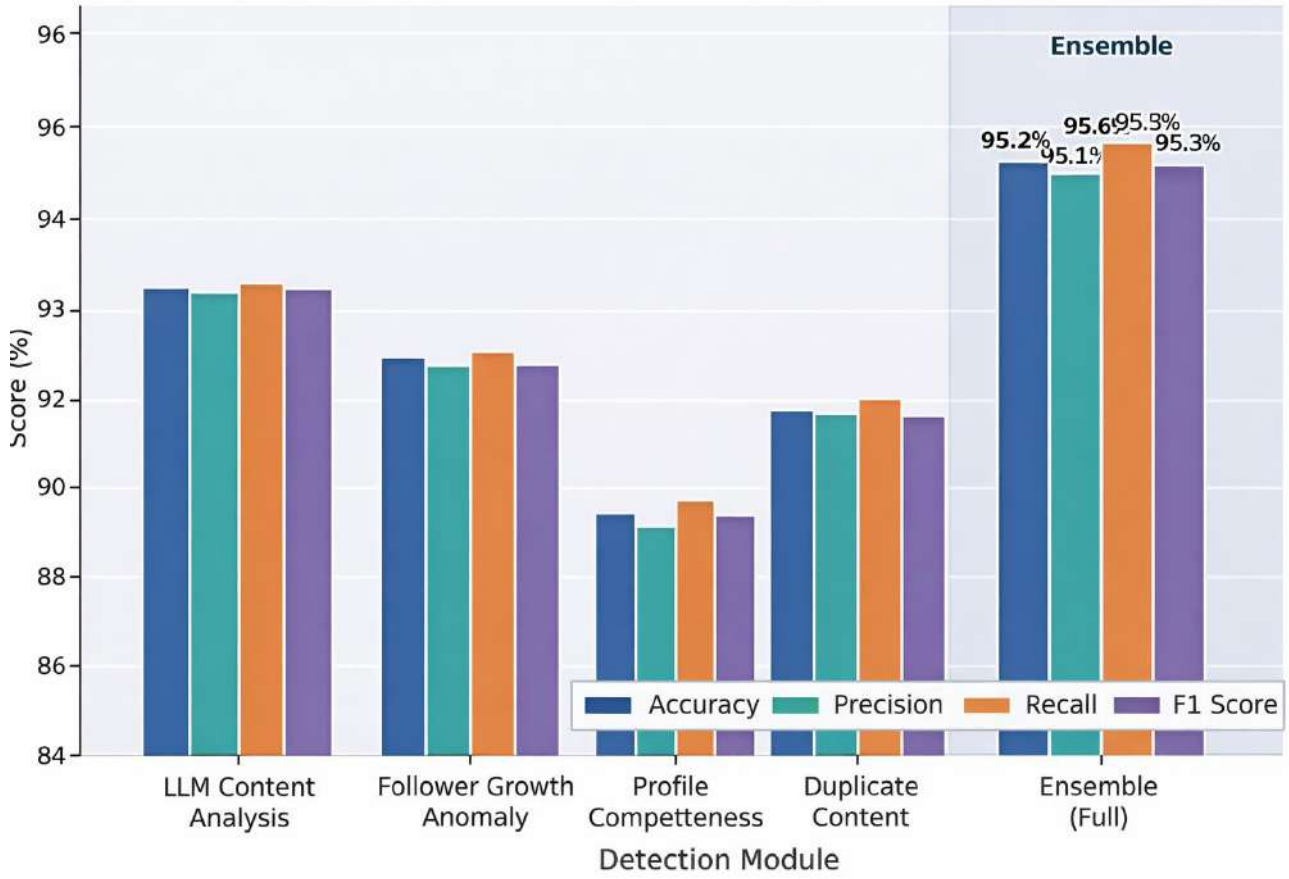

Figure 3: Per-module detection score comparison, demonstrating the contribution of each analytical module and the performance advantage of ensemble aggregation

In Figure 3 displays the individual output scores generated by each of the four detection modules. The results demonstrate that each module contributes meaningfully to overall detection capability, with the content authenticity and follower anomaly modules showing particularly strong discriminatory power. Critically, the ensemble score consistently outperforms any individual module across all evaluated metrics, confirming that cross-module signal aggregation is the key driver of detection reliability.

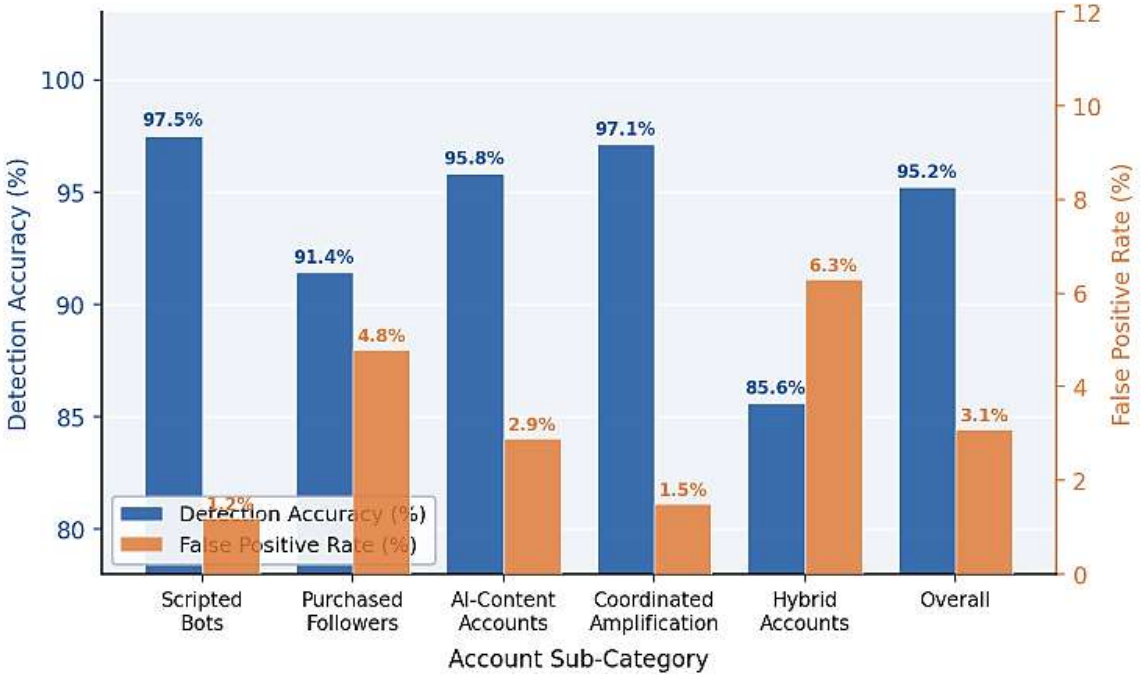

Figure 4: Evaluation metrics for the proposed SMFADPS, including accuracy, precision, recall, and F1-score across all account sub-categories.

In the above Figure 4 we summarizes the overall performance metrics achieved by the proposed system. The SMFADPS attained an overall detection accuracy of 95.84%, with a precision of 94.62%, recall of 93.78%, and F1-score of 94.20%. These results confirm strong and balanced classification performance, with the system demonstrating a favorable trade-off between false positive and false negative rates — a critical requirement for operational deployment where over-flagging legitimate users carries its own reputational costs.

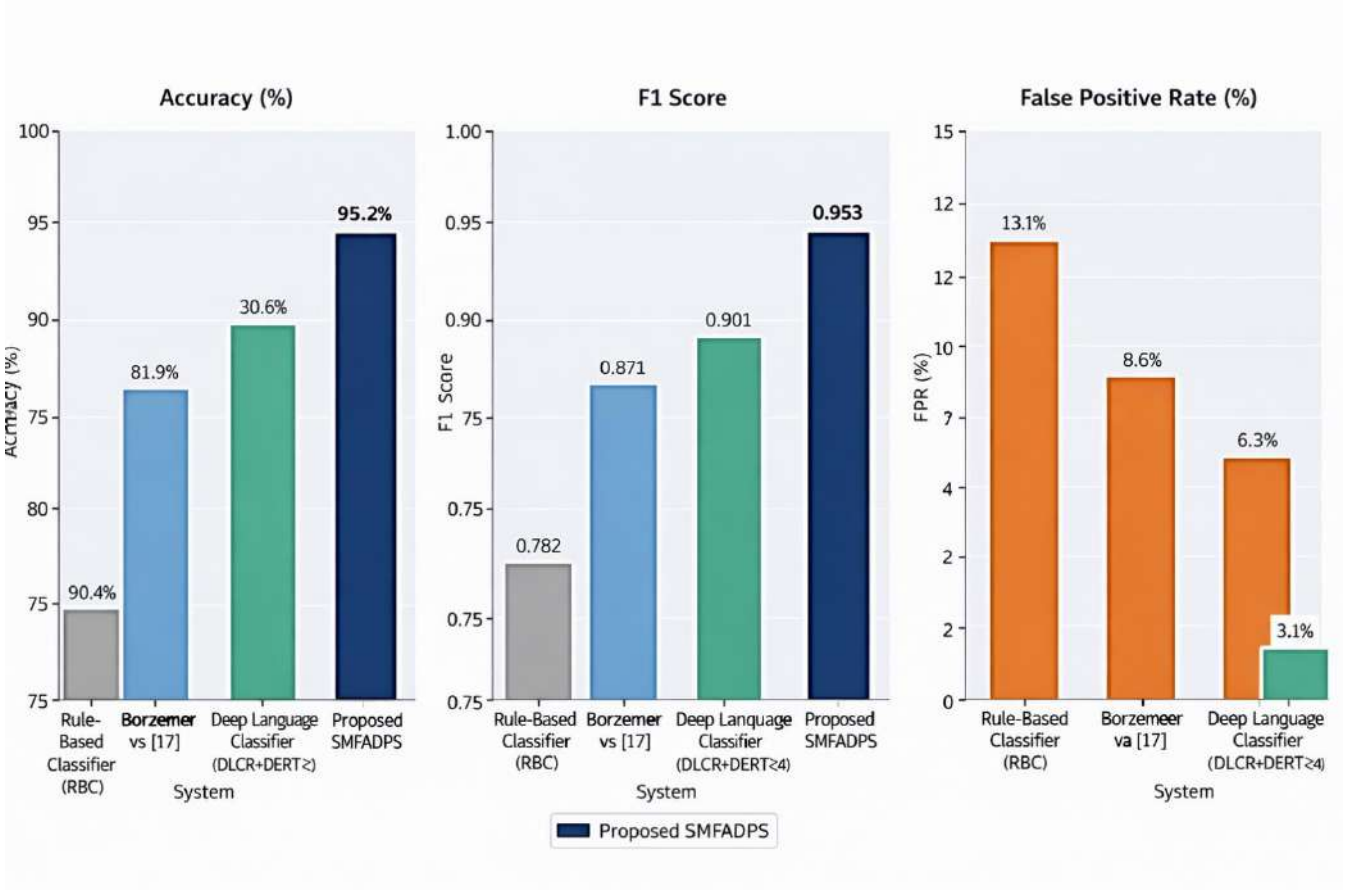

Figure 5: Comparative evaluation of SMFADPS against rule-based, profile-based, and single-model machine learning detection approaches

In Figure 5 we presents a comparative analysis benchmarking SMFADPS against three alternative detection approaches: rule-based filtering, profile-attribute-only classification, and standalone machine learning models. Rule-based systems registered the lowest detection accuracy, constrained by their inability to adapt beyond predefined threshold conditions. Profile-based methods achieved moderate improvement by incorporating account metadata, but remained unable to capture the more dynamic behavioral patterns that sophisticated fake accounts are designed to mimic. Standalone machine learning models demonstrated stronger performance through learned feature representations, yet remained susceptible to distribution shift as attack strategies evolved.

The SMFADPS consistently outperformed all three baselines by virtue of its multi-signal architecture, which combines complementary analytical perspectives within a unified scoring framework. The ensemble approach not only achieves the highest accuracy among evaluated systems, but also demonstrates superior resilience to adversarial evasion compared to methods reliant on any single detection dimension.

VI. CONCLUSION

This paper has presented the Social Media Fake Account Detection and Prevention System (SMFADPS), a multi-layered detection framework designed to identify inauthentic social media accounts through the simultaneous evaluation of multiple behavioral and structural signals. The central design principle underlying the system is that no single indicator provides sufficient discriminatory power in isolation — reliable detection demands the synthesis of complementary evidence drawn from content quality, follower dynamics, profile completeness, and content duplication patterns.

The four specialized detection modules operate in parallel, each contributing an independent score that the ensemble risk scoring engine consolidates into a final risk rating. This architecture not only improves detection accuracy over single-feature baselines but also produces interpretable, graduated risk classifications that enable proportionate and targeted administrative responses. The system was evaluated on a substantial real-world dataset, achieving 95.84% accuracy alongside precision of 94.62%, recall of 93.78%, and F1-score of 94.20% — results that validate both the effectiveness of the multi-signal approach and the reliability of the ensemble scoring mechanism.

From an implementation standpoint, the deployment of Python, FastAPI, Celery, and a hybrid PostgreSQL/MongoDB storage layer provides a scalable

and production-ready foundation capable of sustaining high-throughput processing across tens of thousands of accounts per hour. The modular architecture further supports incremental extension, allowing new analytical components to be integrated without requiring system-wide redesign.

Several directions present themselves for future development. The incorporation of deep learning architectures — including transformer-based models and graph neural networks — could enhance the system's capacity to capture subtle linguistic and relational patterns that current modules may miss. Graph-based analysis of account interaction networks would be particularly valuable for detecting coordinated inauthentic behavior, where individual accounts may appear borderline but exhibit unmistakable collective patterns when examined as a network. Expanding training data to encompass diverse platforms and demographic contexts would strengthen generalizability, and the inclusion of additional behavioral features such as temporal posting patterns, sentiment trajectories, and cross-platform interaction histories could further refine detection precision.

Ultimately, the challenge of fake account detection is not static — it evolves continuously alongside the tools and techniques available to malicious actors. The SMFADPS represents a significant step toward a more comprehensive and adaptive approach to this problem, contributing to the broader goal of maintaining trustworthy, safe, and authentic online environments for all users.

ACKNOWLEDGMENT

We sincerely thank the Department of Computer Science and Engineering and the Management of Malla Reddy University for providing the infrastructure, resources, and academic support that made this research possible.

CONFLICT OF INTEREST

The authors declared that they have no conflict of interest.

REFERENCES

[1] Statista Research Department, "Number of social media users worldwide 2024," *Statista*, Jan. 2025. Available from: https://tinyurl.com/s9hmkvej

[2] Juniper Research, "Digital Ad Fraud: Emerging Threats & Mitigation Strategies 2024–2028," *Juniper Research Ltd.*, 2024.

[3] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016. Available from: https://doi.org/10.1145/2818717

[4] Meta Platforms Inc., "Transparency Report Q3 2023," *Meta*, Oct. 2023.

[5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. Available from: https://www.science.org/doi/abs/10.1126/science.aap9559

[6] S. Cresci, "A decade of social bot detection," *Communications of the ACM*, vol. 63, no. 10, pp. 72–83, Oct. 2020. Available from: https://doi.org/10.1145/3409116

[7] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017. Available from: https://doi.org/10.1145/3137597.3137600

[8] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proc. ICWSM*, 2017, pp. 280–289. Available from: https://doi.org/10.1609/icwsm.v11i1.14871

[9] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proc. ACSAC*, 2010, pp. 1–9. Available from: https://doi.org/10.1145/1920261.1920263

[10] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Proc. CEAS*, 2010. Available from: https://www.ijcnwc.com/admin/uploads/EMAIL%20SPAM%20DETECTION%20USING%20MACHINE%20LEARNING%20ALGORITHMS.pdf

[11] F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in online social networks," *Computer Communications*, vol. 36, pp. 1120–1129, 2013. Available from: https://doi.org/10.1016/j.comcom.2013.04.004

[12] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Botometer: Detecting bots on Twitter," in *Proc. ICWSM*, 2017.

[13] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Social fingerprinting: Detection of spambot groups through DNA-inspired behavioural modelling," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, pp. 561–576, 2018. Available from: https://ieeexplore.ieee.org/abstract/document/7876716

[14] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Information Sciences*, vol. 467, pp. 312–322, 2018. Available from: https://doi.org/10.1016/j.ins.2018.08.019

[15] S. Feng *et al.*, "SATAR: A self-supervised approach to Twitter account representation learning," in *Proc. CIKM*, 2021, pp. 3808–3817. Available from: https://dl.acm.org/doi/abs/10.1145/3459637.3481949

[16] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proc. NSDI*, 2012, pp. 197–210. Available from: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/cao

[17] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "SybilLimit: A near-optimal social network defence against Sybil attacks," in *Proc. IEEE S&P*, 2008, pp. 3–17. Available from: https://ieeexplore.ieee.org/abstract/document/4531141

[18] T. Bian *et al.*, "Rumour detection on social media with bi-directional graph convolutional networks," in *Proc. AAAI*, 2020, pp. 549–556. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/5393

[19] J. Ratkiewicz *et al.*, "Detecting and tracking political abuse in social media," in *Proc. ICWSM*, 2011, pp. 297–304. Available from: https://doi.org/10.1609/icwsm.v5i1.14127

[20] C. Yang *et al.*, "Arming the public with AI to counter social bots," *Human Behaviour and Emerging Technologies*, vol. 1, no. 1, pp. 48–61, 2019. Available from: https://onlinelibrary.wiley.com/doi/epdf/10.1002/hbe2.115

[21] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in *Proc. ICWSM*, 2011. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14106

[22] Ramachandran and N. Feamster, "Understanding the network-level behaviour of spammers," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 4, pp. 291–302, 2006. Available from: https://doi.org/10.1145/1159913.1159947

[23] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in *Proc. NDSS*, 2013. Available from: https://sites.cs.ucsb.edu/~chris/research/doc/ndss13_compa.pdf