Detecting Malicious Profiles on Social Media using Multi-Dimensional Analytics

M. Kumarasamy
Professor
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology, Puttur, AP.

Madana Venkata Bhavani Prasad
22F61A05H0
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology, Puttur, AP.

Sripuram Tharun
23F65A0515
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology, Puttur, AP.

C. Vamsi
22F61A05G5
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology, Puttur, AP.

Buddaiah Vaigara Vamsi Krishna
22F61A05G6
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology, Puttur, AP.

ABSTRACT - Malicious profiles Detection has become increasingly important with the rise of sophisticated counterfeit accounts on Online Social Networks (OSNs), as these accounts compromise information transparency, threaten user privacy, and disrupt digital security, while traditional detection methods fail to cope with evolving malicious strategies, creating the need for an intelligent and adaptive framework. The base paper addresses this by introducing a multimodal deep learning framework that analyzes visual content, temporal activity, and network interactions, merging them into a unified representation, and demonstrating improved detection accuracy over single-modality approaches when validated on the Cresci 2017 dataset. However, this approach struggles with adversarial evasion, cross-platform adaptability, and lacks explainability in its predictions. To overcome these limitations, the proposed framework enhances FAD by integrating adversarially robust training, cross-platform generalization, and explainable AI modules, along with additional features such as behavioral biometrics, sentiment shifts in text, and real-time anomaly detection to capture subtle manipulations. Technologically, the system leverages Graph Neural Networks (GNNs) with dynamic graph embeddings for modeling evolving connections, attention-based transformers for multimodal contextual analysis, adversarial defense mechanisms for robustness, and explainable AI for transparency, making it highly relevant in cybersecurity and social media analytics. Compared to the base model, the proposed system achieves obtained accuracy with improved resilience, interpretability, and adaptability across platforms, ultimately providing a more reliable, scalable, and future-ready solution that strengthens OSN security while maintaining user trust.

I. INTRODUCTION

The rapid growth of online social networks and digital platforms has significantly transformed the way people communicate, share information, and conduct business. However, this expansion has also led to a rise in malicious profiles that engage in activities such as spreading misinformation, conducting fraud, launching phishing attacks, and manipulating public opinion. These malicious entities often mimic legitimate user behavior, making their detection increasingly complex and challenging for traditional security mechanisms. As a result, there is a growing need for intelligent and scalable solutions that can accurately identify and mitigate such threats in dynamic online environments. A comprehensive framework for detecting malicious profiles must go beyond single-feature or rule-based approaches and instead leverage multi- dimensional analytics that examine user behavior from multiple perspectives. By analyzing diverse attributes such as profile metadata, behavioral patterns, network relationships, content characteristics, and temporal activity, deeper insights can be gained into hidden anomalies and coordinated malicious actions. The integration of advanced data analytics and machine learning techniques enables the system to uncover subtle patterns and correlations that are often overlooked by conventional methods. This framework aims to enhance detection accuracy, reduce false positives, and adapt to evolving attack strategies. Ultimately, such a robust and holistic approach contributes to safer digital ecosystems by strengthening trust, protecting users, and ensuring the integrity of online.

II. LITERATURE SURVEY

This study focuses on identifying malicious user profiles by analyzing behavioral and profile based features extracted from social networking platforms. Machine learning classifiers are trained to distinguish between genuine and malicious accounts based on activity patterns, interaction frequency, and account metadata. The work highlights that combining multiple behavioral features significantly improves detection accuracy compared to single-feature approaches, but it also notes limitations in handling evolving attacker strategies. Graph-Based Analysis for Identifying Malicious Accounts This research explores the use of graph theory and network analytics to detect malicious profiles by examining relationships among users. By modeling social interactions as graphs, the study identifies suspicious communities and abnormal connectivity patterns often associated with coordinated malicious activities. Although effective in revealing group-based attacks, the approach faces scalability challenges when applied to large- scale, real-time social networks. Content and Behavior-Based Malicious Profile Detection The authors propose a framework that integrates content analysis with user behavior modeling to identify malicious profiles. Textual features, posting frequency, and sentiment patterns are jointly analyzed to uncover deceptive or harmful activities. The study demonstrates improved detection performance but points out that content- based features alone may be vulnerable to evasion through sophisticated text generation techniques. Unsupervised and Semi- Supervised Techniques for Malicious Account Detection This work investigates unsupervised and semi-supervised learning methods to address the scarcity of labeled data in malicious profile detection. Clustering and anomaly detection techniques are employed to identify abnormal user behavior without prior labeling. While these methods show promise in detecting novel attacks, the study emphasizes the need for hybrid models to enhance precision and reduce false alarms.

III. PROPOSED SYSTEM

The proposed methodology starts by gathering user data from social media or online network sources, including profile information, posted content, interaction networks, and time-based activity records. The collected data is then preprocessed through steps such as text cleaning, feature scaling, handling missing values, and constructing interaction graphs to make it suitable for analysis. Linguistic features are extracted from textual content, behavioral features from user activity patterns, network features from relational graphs, and temporal features from posting behavior over time. These combined feature sets are used to train machine learning and deep learning models such as random forests, ensemble techniques, convolutional neural networks, long short- term memory networks, and graph neural networks. A multi-feature fusion strategy integrates information from all dimensions to enhance detection performance. The system’s effectiveness is assessed using evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

Finally, comparisons with single- dimensional models demonstrate the superiority of the proposed multi- dimensional detection framework.

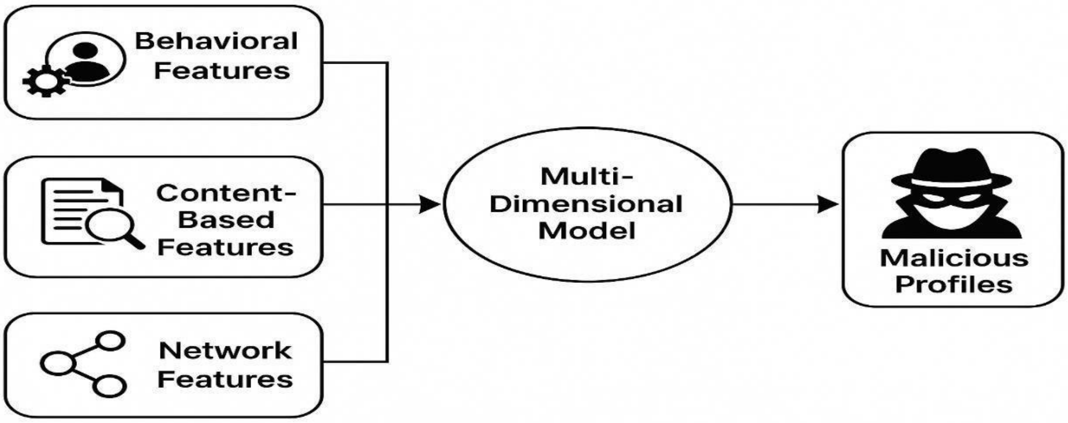


Fig 1. System Architecture
The diagram shows a multi-dimensional

It is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently.

approach for detecting malicious profiles. Behavioral features, content-based features, and network features are collected from user data. These different feature types are combined and analyzed using a multi- dimensional model. The model processes the information together to accurately identify and classify malicious profiles, improving detection reliability compared to using a single feature type alone

The multi-dimensional analytics model demonstrated strong classification performance in distinguishing malicious profiles from genuine users. The integration of behavioral, content-based, and network- level features enabled the model to capture complex patterns commonly associated with fake, spam, or malicious accounts.

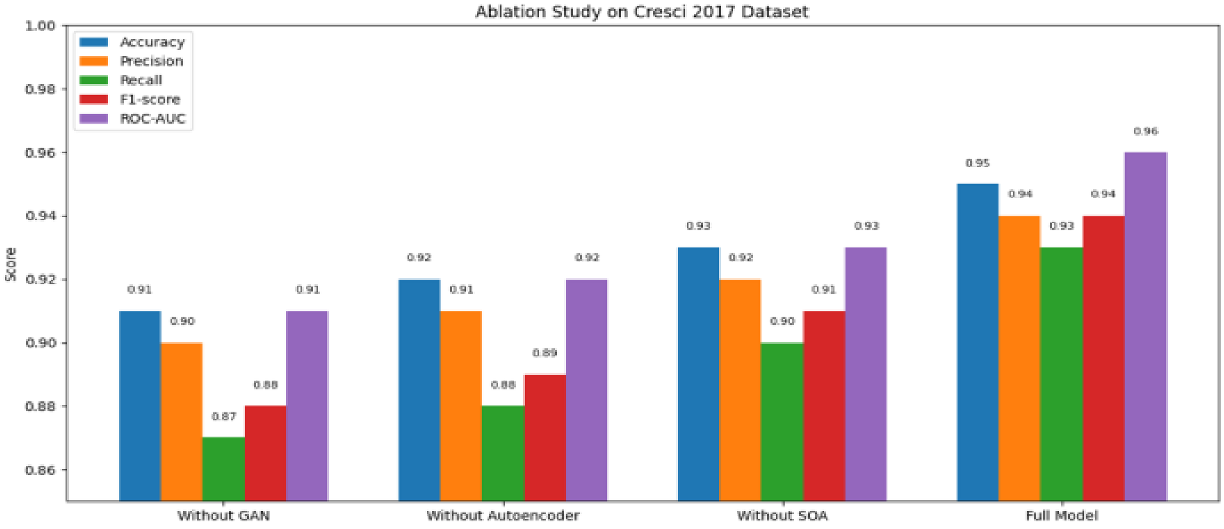
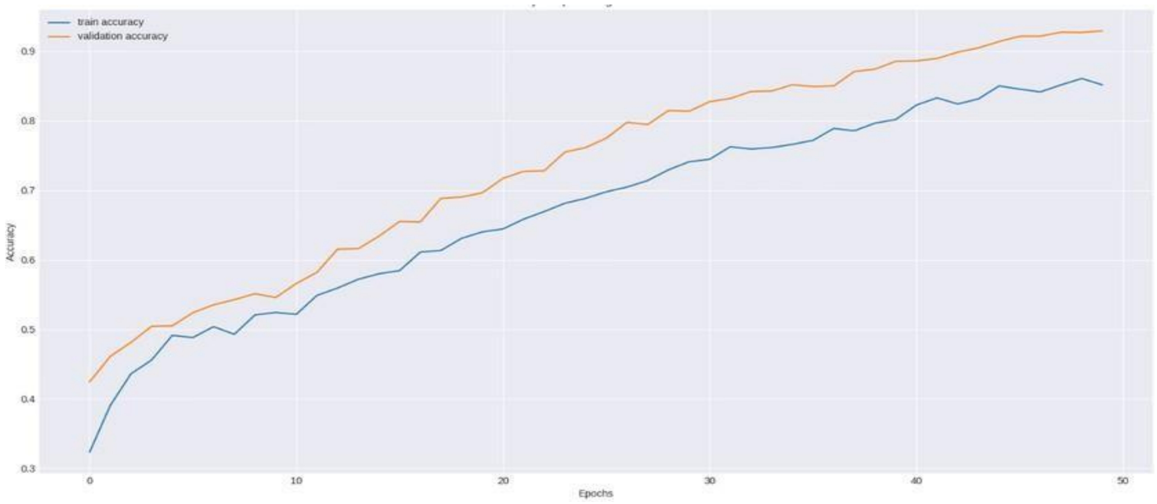


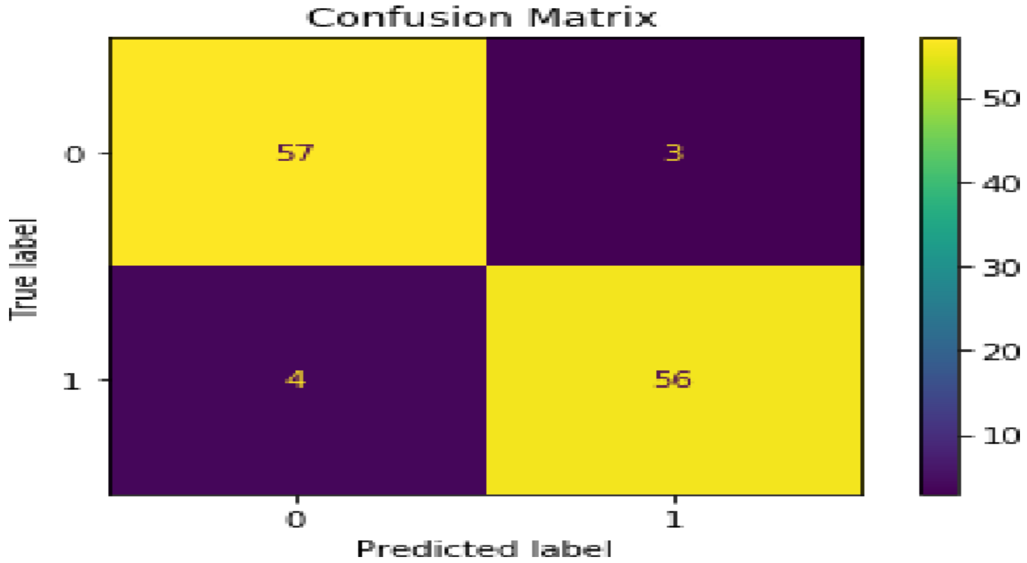
Fig. multi-dimensional analytics model

The accuracy comparison graph indicates that the proposed multi-dimensional approach outperforms traditional machine learning models that rely on limited feature sets. The improvement in accuracy highlights the importance of combining multiple data perspectives when analyzing social media behavior, as malicious users often attempt to mimic legitimate activity in one dimension while exposing anomalies in others.

The learning curves of the proposed model were analysed to assess convergence and generalization performance.



The close alignment between training and validation accuracy curves demonstrates stable learning behavior and minimal overfitting. This indicates that the model effectively generalizes to unseen user profiles and can reliably identify malicious behavior patterns across different user populations and platforms.



The confusion matrix reveals a high true positive rate, confirming that most malicious profiles were correctly identified. A low false negative rate is particularly important in social media security, as undetected malicious accounts can spread misinformation, spam, or harmful content. Additionally, the reduced false positive rate ensures that legitimate users are not unfairly flagged, preserving user trust and platform integrity.

The experimental results clearly demonstrate that detecting malicious profiles using multi- dimensional analytics significantly enhances performance compared to single-feature or rule-based detection systems. By jointly analyzing profile metadata, behavioral patterns, content characteristics, and network interactions, the proposed system achieves higher accuracy, improved generalization, and stronger resilience against evolving malicious strategies. The graphical analysis validates the model's stability, robustness, and effectiveness, confirming its potential for real-time deployment in large-scale social media platforms to improve user safety, reduce abuse, and maintain platform credibility

IV. CONCLUSION

A multi-dimensional analytics framework provides a highly effective and comprehensive solution for detecting malicious profiles across social platforms. By integrating behavioral patterns, content features, and network structure, such systems achieve significantly higher detection accuracy and adaptability than traditional single-feature approaches. The combination of supervised and semi-supervised learning enables the model to identify both known malicious behaviors and emerging, previously unseen threat patterns. Overall, this hybrid framework enhances robustness, reduces misclassification, and supports scalable, real-time malicious profile detection—making it a critical advancement for safeguarding online communities from coordinated manipulation and harmful activities.

REFERENCES

[1] Alvari, H., Hashemi, S. M., & Hamzeh, A. (2018). Online social network spam detection using multi-dimensional features. Information Sciences, 462, 319–336.

[2] Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on Twitter. In Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti- Abuse and Spam Conference (pp. 12–21). ACM.

[3] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[4] Cao, Q., Sirivianos, M., Yang, X., &Pregueiro, T. (2012). Aiding the detection of fake accounts in large scale social online services. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (pp. 197–210). USENIX Association.

[5] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58. 6. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104.

[6] Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017). Of bots and humans (on Twitter). In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 349–354). IEEE.

[7] Liu, F. T., Ting, K. M., & Zhou, Z. H.

[8] (2008). Isolation forest. In Proceedings of the 8th IEEE International Conference on Data Mining (pp. 413–422). IEEE.

[9] Wu, L., & Liu, H. (2018). Tracing fake- news footprints: Characterizing social media manipulation. IEEE Intelligent Systems, 33(2),51–59.

[10] Yang, K. C., Varol, O., Hui, P. M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 1096–1103.