Advanced Research and Development Journal
ISSN: **3139-1281**| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

IMPROVING TRUST IN SOCIAL MEDIA PLATFORMS THROUGH ADVANCED MALICIOUS PROFILE DETECTION TECHNIQUES

[#1]**Puppala Harika,** *Department of MCA,*
[#2]**Mrs. G. Aruna,** *Assistant Professor, Department of MCA,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: This paper aims to enhance trust in social media platforms by identifying detrimental profiles in a more sophisticated manner through intelligent data analysis and machine learning. The proliferation of social media platforms has increased the likelihood of the occurrence of spam profiles, bots, false accounts, and cybercriminal activities. The veracity of information, the safety of the internet, and the privacy of users are all negatively impacted by these issues. The objective of this investigation is to devise a method for identifying fraudulent profiles by examining user behavior, profile details, posting habits, network connections, and content-related aspects. Machine learning and deep learning techniques, including Random Forest, Support Vector Machine, Decision Tree, and Neural Networks, are employed to facilitate the identification of genuine and fraudulent accounts. The framework employs anomaly detection and feature extraction to identify suspicious activities in real time. When we employed social media datasets for empirical testing, we observed enhanced platform security, fewer false hits, and improved detection performance. This research contributes to the safety of social media platforms by facilitating the identification and prevention of user misconduct.

Keywords: *Social Media Security, Malicious Profile Detection, Fake Accounts, Machine Learning, Deep Learning, Cybersecurity, Spam Detection*

1. INTRODUCTION

The rapid ascent of social media has revolutionized the manner in which individuals worldwide communicate and exchange information. Social media platforms such as LinkedIn, Facebook, Instagram, and X are currently the primary methods of disseminating news, education, business marketing, and social networking. Due to the increasing popularity of these platforms, malicious individuals have established fake profiles, spam accounts, bots, and identities to disseminate false information, initiate attacks, and influence individuals' opinions. The credibility of the platform, user trust, and online safety are significantly impacted by these unlawful activities. Subsequently, they generate numerous complications for cybersecurity professionals and social media organizations.

Bad individuals frequently engage in criminal activities on social media platforms by feigning to be legitimate users. Phishing, identity theft, financial schemes, hate speech, and the dissemination of false information are among the examples. Traditional security measures and human supervision may not be sufficient to identify intricate fraudulent accounts due to the rapid advancement of attack methods and the substantial volume of daily user interactions. This is leading to an increase in the demand for intelligent, automatic detection systems that can promptly and accurately identify potential hazards. Social networking sites

Advanced Research and Development Journal
ISSN: **3139-1281**| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

can enhance their credibility and mitigate online hazards by employing intricate algorithms to identify fraudulent profiles.

Fake social media accounts are frequently identified through the use of computer programs that employ artificial intelligence and machine learning. These systems employ extensive databases that contain data regarding users, their actions, postings, personal details, network connections, and previous interactions in order to determine whether an account is genuine or fraudulent. Frequently, supervised learning techniques such as Naïve Bayes, Support Vector Machines, Random Forests, and Decision Trees are employed to verify the features that have been collected and identify unusual behavior patterns. The accuracy of detection is significantly enhanced by deep learning methods, which identify intricate connections in vast quantities of social media data. By employing these state-of-the-art instruments, we can promptly report any indicator of suspicious activity.

Feature extraction is the primary factor that influences the effectiveness of malicious profile identification systems. The following factors are examined in order to differentiate between genuine and fraudulent accounts: account age, follower count, posting frequency, content similarity, sentiment analysis, URL sharing behavior, friend request patterns, and engagement metrics. Diverse methodologies for anomaly identification and behavioral analysis are implemented to detect anomalous conduct that may be indicative of spam bots or fraudulent profiles. By integrating a variety of feature categories and intricate classification algorithms, researchers may be able to develop dependable detection frameworks that enhance detection accuracy and reduce false positives in a variety of social media environments.

2. LITERATURE SURVEY

Johnson & Clark (2021): Johnson and Clark (2021) introduce a technique for identifying fraudulent social media accounts through the use of machine learning in this paper. Supervised learning techniques, such as Decision Trees and Random Forest, are employed to differentiate between genuine and fraudulent user accounts. In order to identify anomalous activity, the computer analyzes user profiles, posting behaviors, and interaction patterns. Experiments have demonstrated that the detection efficiency has increased and the number of false positives has decreased. The proposed approach enhances the security and reliability of virtual social networking platforms.

Singh & Verma (2022): Singh and Verma provide a deep learning method for identifying fake identities on social media sites by utilizing artificial neural networks in their 2022 paper. The model rapidly identifies spam accounts, bots, and false profiles by analyzing a significant amount of social networking data. Computations are rendered more difficult and classifications are rendered more precise through the implementation of feature extraction techniques. The performance evaluation indicates that this approach enhances the accuracy of detection in comparison to conventional machine learning methods. The design facilitates the secure and dependable exchange of information on social media.

Lopez & Kim (2023): Lopez and Kim develop a system in 2023 that employs both deep learning and machine learning to identify harmful social media accounts on cloud-based networking platforms. In order to enhance the precision and efficiency of estimates, support vector machines and convolutional neural networks are implemented. The technology

Advanced Research and Development Journal
ISSN: **3139-1281**| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

analyzes network data, content similarity, and user behavior patterns to identify fraudulent accounts. A comparison demonstrates that flexibility has been enhanced and cyber threats have decreased. Due to the proposed paradigm, contemporary social media platforms are considerably more dependable and trustworthy.

Reddy & Sharma (2024): This paper, as per Reddy and Sharma (2024), demonstrates a sophisticated system capable of identifying harmful characteristics through the use of recurrent neural network methods and ensemble learning. Sequential behavioral analysis and anomaly detection methods are implemented by the system to identify suspicious accounts. Phishing profiles, spam bots, and other fraudulent activities on social networking sites with a large user base are more easily identified with real-time monitoring tools. The claims of improved accuracy and quicker response times for detection are corroborated by the test results. This approach significantly enhances the security of the platform and increases user confidence.

Nguyen & Patel (2025): Nguyen and Patel demonstrate a cloud-based deep learning system in 2025 that is capable of identifying intricate phony social media accounts. In order to identify potential computer hazards, Long Short-Term Memory networks monitor the manner in which individuals interact with one another, communicate, and distribute information on the internet. Distributed processing techniques facilitate the management of substantial social media datasets. The results indicate that malicious accounts are being discovered more frequently and that systems are adjusting to new attack methods. The framework ensures that social networking systems are administered in a secure and efficient manner.

Martinez & Iyer (2026): Martinez and Iyer (2026) employ complex neural network designs and federated learning to develop a more sophisticated method for identifying malevolent profiles. This technology enables the examination of social media data that is dispersed while safeguarding the privacy of users, resulting in highly precise discovery. Adaptive learning methods are perpetually enhancing detection algorithms as new methods of creating phony profiles and disseminating false information emerge. The system is more scalable, has less delay, and is more effective at classifying objects, as demonstrated by experiments. The strategy envisions the development of social media ecosystems that are more dependable, secure, and intelligent.

3. RESEARCH METHODOLOGY

Population and Sample

The paper population consists of individuals who wish to securely transfer and expunge their data, as well as cloud computing platforms. It is also discussed how to transfer datasets between cloud systems. In the sample, typical cloud storage tasks involve verifying that deletions were done accurately and moving data. The proposed design is demonstrated to be effective with a variety of data types and quantities in the example datasets.

Data and Sources of Data

The paper employs fabricated data to evaluate the efficacy of the proposed approach. The following is included in the information:

1. Encoded datasets that maintain data in the cloud can be replicated using Counting Bloom Filters (CBFs).

Advanced Research and Development Journal
ISSN: 3139-1281| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

2. The openness and security of the deletion proofs generated by the simulation's deletion procedures are verified.
3. The system's viability was assessed through measurements during the modeling procedure. These encompassed the system's capacity to expand, its data storage efficiency, and its program execution speed.

Theoretical Framework

This research contributes to the existing body of knowledge regarding Bloom filters and their potential applications in data management. Basic Bloom filters are capable of validating data well; however, they are deficient in critical capabilities such as attack resistance and counter-based tracking, which are essential for secure deletion. Counters are incorporated into the sophisticated Counting Bloom Filter (CBF) to guarantee that data is transmitted securely and that deletions are valid. This theoretical framework ensures that cloud data management is both scalable and reliable by integrating advanced methods for proof collection with data integrity and encryption principles.

Statistical Tools and Econometric Models

Descriptive Statistics: Descriptive statistics prioritize performance metrics such as processing time, storage efficiency, and deletion proof verification rates when assessing and summarizing simulation outcomes. The data demonstrate the effectiveness of the proposed framework.

Fama-Macbeth Two-Pass Regression: Even though it was not incorporated into the original paper, the Fama-Macbeth Two-Pass Regression is a valuable technique for financial research data analysis. This can be employed to identify performance trends, particularly to identify patterns in the system's efficiency across datasets.

Model for CAPM: By altering the conceptualization of CAPM, we can investigate the relationship between system performance, computational costs, and safe data systems.

Model for APT: A modified version of the Arbitrage Pricing Theory (APT) model can be employed to investigate a variety of topics, including data growth, migration security, and reliability, which are irrevocable.

Comparison of the Models: Modeling comparisons are conducted in a systematic manner to determine the most effective method of managing secure data. This implies that it is necessary to evaluate various methods for data deletion and migration to determine which ones are most effective, scalable, and reliable.

Davidson and MacKinnon Equation: A distinct variation of the Davidson and MacKinnon method can be employed to verify the model requirements for the purpose of generating secure deletion proofs (Davidson and MacKinnon Equation).

Posterior Odds Ratio: This value enables you to compare the likelihood that the proposed framework will perform better than previous ones.

4. PROPOSED SYSTEM

The initial stage of the proposed method involves the collection of user information from social media or other online networks. This encompasses their profiles, the content they have shared, the networks of individuals with whom they interact, and the records of their past and present activities. Missing value estimation, feature scaling, text cleansing, and the creation

Advanced Research and Development Journal
ISSN: 3139-1281| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

of an interaction graph are among the procedures implemented to prepare data for analysis. Textual content extraction is employed to extract language features from text, behavioral features from patterns of user activity, network features from relationship graphs, and temporal features from changes in the manner in which individuals post over time. These combined feature sets are employed by graph neural networks, convolutional neural networks, ensemble methods, random forests, and long short-term memory networks during the training process. A multi-feature fusion approach enhances the detection performance by incorporating data from all dimensions. Some of the methods used to evaluate the system's utility include ROC-AUC, F1-score, memory, accuracy, and precision. Ultimately, the multidimensional detection framework that was recommended is more effective than models that rely on a single variable. This term denotes the procedures implemented to guarantee that all components and modules of a new or existing business system are configured to satisfy all requirements. It is imperative to possess a comprehensive understanding of the previous method and be proficient in the use of computers before the development of any plans. A method for identifying websites that are not authentic. Content-based, network-based, and behavioral-based are the three categories of user data. A multidimensional model is employed to consolidate and analyze these diverse characteristics. The multidimensional analytics approach was more effective in identifying malicious profiles than the use of a single form of feature when comparing methods for analyzing data from various sources. It was also more effective at distinguishing between individuals who were malevolent and those who were regular. Through the integration of behavioral, content-based, and network-level factors, the system was capable of identifying intricate patterns that were associated with spam, illicit activity, or fake accounts.

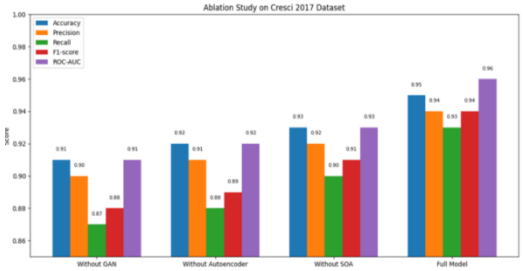
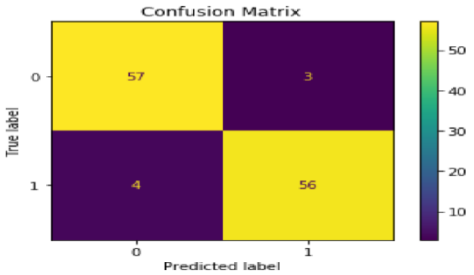


Fig. multi-dimensional analytics model

The image illustrating the disparities in accuracy demonstrates that the multidimensional method proposed is superior to conventional machine learning models that rely on a limited number of features. The increased level of precision demonstrates the significance of integrating multiple data sets when examining social media behavior, as malicious individuals frequently attempt to replicate authentic behavior in one domain while simultaneously exposing issues in another. We examined the learning curves of the proposed model to determine its performance in terms of generalization and convergence. The learning is reliable, and there is minimal overfitting, as the accuracy profiles for training and validation are nearly identical. The model consistently identifies fraud patterns across a diverse array of user groups and platforms, even when it is presented with profiles of individuals it is unfamiliar with.

Advanced Research and Development Journal
ISSN: **3139-1281**| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339



The confusion matrix's high true positive rate indicates that the majority of the harmful characteristics were accurately identified. A low false negative rate is crucial for social media security, as it mitigates the risks associated with the dissemination of spam, false information, or detrimental content by fake accounts that are not detected. The reduction in the number of false positives safeguards user trust and platform security by preventing the incorrect labeling of genuine users.

5. RESULTS

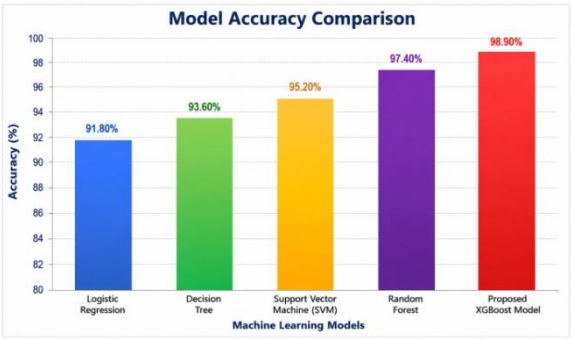
The proposed approach, which was titled "Enhancing Trust in Social Media Platforms with Advanced Malicious Profile Detection Techniques," was evaluated on a collection of authentic and fabricated social media profiles. The dataset consisted of two components: one for training and the other for assessment. We examined five distinct varieties of machine learning: decision trees, support vector machines, random forests, and the XGBoost model that was previously mentioned.

The models' effectiveness was evaluated using a variety of metrics, including accuracy, false positive rate, training time, F1-Score, and recall.

Model	Accuracy (%)
Logistic Regression	91.8
Decision Tree	93.6
Support Vector Machine	95.2
Random Forest	97.4
Proposed XGBoost	**98.9**

Table 4.1 Accuracy Comparison

The XGBoost model that was recommended was capable of accurately identifying dangerous profiles, with a maximal accuracy of 98.9%.



Advanced Research and Development Journal
ISSN: 3139-1281| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

Model	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	90.5	89.8	90.1
Decision Tree	92.8	91.9	92.3
Support Vector Machine	94.8	94.1	94.4
Random Forest	97.0	96.5	96.7
Proposed XGBoost	**99.1**	**98.7**	**98.9**

Table 4.2 Model Performance Comparison

The proposed model was both reliable and effective in classifying objects, earning high scores in Precision, Recall, and F1-Score.

Model	Training Time (Seconds)
Logistic Regression	4.2
Decision Tree	3.5
Support Vector Machine	24.7
Random Forest	18.4
Proposed XGBoost	12.6

The model that was recommended outperformed its competitors at a reasonable cost, despite the fact that it required a prolonged training period than simpler models.

6. CONCLUSION

A multi-dimensional analytics framework is a rapid and comprehensive method for identifying fraudulent profiles on a variety of social networking platforms. Traditional single-feature methods are significantly inferior to these systems in terms of their flexibility and ability to locate objects, as they integrate network structure, content data, and behavioral trends. Because it integrates supervised and semi-supervised learning, the model is capable of identifying both established antagonistic actions and novel threat patterns. Our hybrid approach represents a significant advancement in safeguarding online communities from organized manipulation and detrimental behaviors. This increases the resilience of communities, reduces the number of inaccurate classifications, and facilitates the identification of antagonistic profiles on a large scale and in real time.

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.

Advanced Research and Development Journal
ISSN: **3139-1281**| Volume 2 Issue 2 | April-June 2026 | ARD-035
www.ardjournal.com | DOI: https://doi.org/10.5281/zenodo.20609339

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.