



FAKE PROFILE DETECTION ON SOCIAL MEDIA

Ms.Hole.S.S¹, Nikat Samruddhi²

¹ Faculty Guide & Supervisor | ² Dattakala Group Of Institutions Faculty Of Engineering Swami - Chincholi tal. Daund dist. Pune, India

Corresponding author: nikatsamruddhi@gmail.com

ARTICLE ID IJREK-2026-00161	ARTICLE TYPE Research Article	RESEARCH AREA Machine Learning & Deep Learning	PUBLICATION Volume 1, Issue Issue 1 (2026)
RECEIVED 25 October 2026	REVISED 27 October 2026	ACCEPTED 26 October 2026	PUBLISHED 27 October 2026

Abstract

Social media platforms enable communication, news sharing, and online communities, but fake or inauthentic accounts may be used for spam, impersonation, scams, coordinated manipulation, and misleading content. This paper proposes a machine-learning-based Fake Profile Detection system that evaluates available account and activity features and flags profiles that may require further review. The planned workflow includes dataset selection, data validation, preprocessing, feature engineering, model training, threshold selection, and evaluation. Candidate methods include Logistic Regression, Decision Tree, Random Forest, and anomaly-detection techniques when suitable. Because the prevalence of fake accounts and the reliability of labels can vary by dataset, accuracy alone is insufficient; precision, recall, F1-score, confusion matrices, and precision-recall analysis should be considered. This is a proposed academic prototype: no measured accuracy or operational effectiveness is claimed. Any real deployment should protect user privacy, limit data collection, reduce bias, and provide human review before consequential action.

Keywords: Fake Profile Detection; Social Media; Machine Learning; Classification; Feature Engineering; Anomaly Detection; Precision; Recall. Research-paper draft for academic learning. Dataset details; implementation; evaluation results

I. INTRODUCTION

A. Background / Context Social networking services allow users to create profiles, share content, follow other accounts, and interact with communities. Fake profiles may imitate real people, automate engagement, distribute spam, or support fraudulent and coordinated activity. At large scale, manual review of every account is difficult. Machine learning can help prioritize profiles for investigation by learning patterns from labelled examples or identifying unusual account behavior. The usefulness of such systems depends on the dataset, feature quality, labeling practices, and evaluation design. B. Problem Statement Fake accounts can resemble genuine users, while genuine users may also display unusual activity. A model can appear successful if the dataset is unbalanced or if training and test records overlap. Account behavior also changes over time, and false positives may unfairly affect legitimate users. A transparent prototype is therefore needed to preprocess authorized data, compare suitable models, measure missed fake profiles and incorrect flags, and present predictions as risk indicators rather than proof that an account is fake. C. Aim and Objectives The aim is to design a prototype that identifies social media profiles that may require additional review. Objectives include: • Select a documented and appropriately licensed dataset with a clear definition of fake and genuine accounts. • Clean records and prepare available numerical, categorical, and activity-based features. • Avoid data leakage and address class imbalance using appropriate training-only methods where justified. • Train and compare classification or anomaly-detection models. • Evaluate models using precision, recall, F1-score, confusion matrix, and precision-recall measures. • Display a risk flag with understandable reasons where possible and support human review. • Document privacy, security, bias, platform-policy, and model-drift limitations. D. Significance and Scope The project demonstrates how machine learning can support the review of social media profiles. It is intended for academic experimentation and does not establish a user's identity or prove malicious intent. A prototype should not automatically suspend accounts, expose personal information, or make irreversible decisions without appropriate validation and review.

II. LITERATURE REVIEW / RELATED WORK

Research on fake-account detection explores supervised classification when reliable labels are available and anomaly detection when labels are incomplete. Logistic Regression provides a simple baseline; Decision Trees and Random Forests can model non-linear relationships among profile and activity features. Other approaches may analyze graph structure, posting patterns, account metadata, or text, where those data are lawfully available. Results can be affected by class imbalance, label noise, platform differences, coordinated behavior, and changing account strategies. Research Gap A useful student-level system should clearly document the dataset, target definition, and feature availability; compare a baseline with at least one stronger model; report precision and recall alongside accuracy; and explain how a flagged profile would be reviewed. It should distinguish a model-generated risk indicator from a confirmed fake account and discuss fairness and changing behavior over time.

III. PROPOSED SYSTEM

TABLE I

Component	Main responsibility
Data input	Loads authorized profile records and checks required fields.
Preprocessing	Handles missing values, duplicates, encoding, and scaling when required.
Feature engineering	Derives supported indicators such as profile completeness or activity summaries, where available.
Detection model	Classifies records or identifies unusual patterns using the selected method.
Evaluation module	Calculates metrics on data that was not used for training.
Results dashboard	Shows the model flag, available explanation, and evaluation summary.
Privacy and audit layer	Limits access to sensitive data and records appropriate system actions.

IV. METHODOLOGY

A. Dataset and Input Features Use a documented, appropriately licensed dataset and record its source, version, label definition, and limitations. Depending on the dataset

and platform permissions, possible features may include account age, profile completeness, follower/following counts, posting frequency, engagement summaries, or verified labels. Not every dataset contains these fields, and a feature should not be assumed available without checking the data. Avoid direct personal identifiers and comply with the dataset's terms of use.

B. Data Preprocessing

- Check missing values, duplicate rows, invalid types, and inconsistent target labels.
- Encode categorical fields and scale numerical fields when required by the selected algorithm.
- Split data before fitting transformations to prevent data leakage.
- Consider class imbalance and apply resampling only to training data if used.
- Where timestamps or related accounts exist, consider time-based or group-aware splits to reduce leakage.

Keep test data separate until final evaluation.

C. Candidate Algorithms

Logistic Regression can serve as a baseline for labelled classification. Decision Tree models are relatively easy to explain, while Random Forests can capture more complex relationships among features. An anomaly-detection method may be explored when labels are limited, but its output and evaluation differ from supervised classification. The chosen model should be justified by dataset suitability, interpretability, computational cost, and the trade-off between false alerts and missed fake profiles.

D. Evaluation Metrics

Accuracy is the proportion of all records classified correctly, but it can be misleading when fake profiles are rare or the dataset is unbalanced. Precision measures how many flagged profiles are labelled fake; recall measures how many labelled fake profiles are detected. F1-score balances precision and recall, while a confusion matrix summarizes true positives, false positives, true negatives, and false negatives. A precision-recall curve can help assess performance across thresholds. The threshold-selection method should be documented and must not be chosen using the final test set.

E. Ethical, Privacy, and Security Considerations

- Use only data the project is authorized to process and respect platform terms.
- Avoid collecting unnecessary personal information or private messages.
- Restrict access to profile records and protect stored data.
- Treat predictions as review alerts, not proof of deception or malicious intent.
- Check false positives and false negatives because both can cause harm.
- Assess whether features create unfair outcomes for particular user groups.
- Monitor changes in account behavior and data distributions before considering deployment.

V. SYSTEM DESIGN AND IMPLEMENTATION PLAN

TABLE II

Layer	Possible student-level choice
Programming language	Python.
Data processing	pandas and NumPy.
Machine learning	scikit-learn: Logistic Regression, Decision Tree, Random Forest; anomaly detection if appropriate.
Visualization	Matplotlib or a simple dashboard.
Interface	Streamlit, Flask, or another guide-approved interface.
Dataset	A documented public dataset suitable for fake-profile classification.

VI. RESULTS AND EVALUATION PLAN

TABLE III

Test ID	Test case	Expected result
T1	Valid profile record is provided.	Input is validated and a model output is returned.
T2	A required field is missing.	System requests correction or handles the missing field safely.
T3	A field has an invalid type or range.	A clear validation message is displayed.
T4	Model is evaluated on held-out test data.	Metrics are calculated from test predictions.
T5	A profile is flagged as suspicious.	The interface describes it as a review alert, not a confirmed fake profile.
T6	Unauthorized access to profile records is attempted.	Access is denied.

VII. ADVANTAGES AND LIMITATIONS

TABLE IV

Potential advantages	Limitations
Can prioritize profiles for further review.	False positives may affect genuine users.
Allows comparison of several algorithms.	Reliable labels may be limited or noisy.
Can process large datasets in batch mode.	Model quality depends on data quality and feature availability.
Supports learning about classification and anomaly detection.	Account behavior and platform patterns change over time.
Can be extended with monitoring and dashboards.	A student prototype is not validated for production use.

VIII. APPLICATIONS

- Academic demonstrations of social media profile classification.
- Research on supervised fake-account detection and anomaly detection.
- Educational dashboards for reviewing model metrics.
- Batch analysis of authorized sample profile records.
- Prototype alert systems that route potentially suspicious profiles to a human reviewer.
- Exploration of profile and activity patterns for platform-safety research, where data use is authorized.

IX. FUTURE SCOPE

- Evaluate models on larger, more representative, and more recent datasets.
- Explore time-aware validation and concept-drift monitoring.
- Add interpretable explanations showing which available features influenced a prediction.
- Compare class-weighting, threshold tuning, and training-only resampling methods.
- Improve secure data handling, access control, and audit trails.
- Study false-positive rates across relevant groups where lawful, ethical, and feasible.
- Explore graph-based signals only when reliable relationship data are available and permitted.
- Conduct independent validation and human-impact assessment before any real deployment.

X. CONCLUSION

This paper proposes a machine-learning-based Fake Profile Detection prototype for identifying social media accounts that may require further review. The planned system includes data preprocessing, feature preparation, model training, evaluation, and a dashboard for displaying risk alerts. Logistic Regression, Decision Tree, Random Forest, and suitable anomaly-detection approaches are possible candidates depending on the dataset and label availability. Because dataset labels may be noisy and class distributions may be uneven, evaluation should emphasize precision, recall, F1-score, confusion matrices, and precision-recall analysis rather than accuracy alone. No measured results are claimed in this draft. Future work should prioritize privacy, fairness, human review, and monitoring for changing account behavior.

REFERENCES

- [1] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017. Verify bibliographic details before submission.
- [2] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for Sale: Efficient Detection of Fake Twitter Followers," Decision Support Systems, vol. 80, pp. 56–71, 2015. Verify complete publication details before submission.
- [3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," in Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), 2010. Verify complete bibliographic details before submission.
- [4] scikit-learn developers, "Classification metrics" and "Outlier detection," scikit-learn User Guide. Verify the current documentation URLs and access date.
- [5] Add recent peer-reviewed studies relevant to the exact dataset, platform, selected algorithms, and evaluation approach. Verify all authors, titles, publication details, and DOI before submission. Note: Confirm references, dataset licensing, and the final implementation with the project guide before academic submission.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on <https://www.ijrek.org> | Article ID: IJREK-2026-00161