

Fake Social Media Accounts and Their Detection

Miss. Gayatri R. Gatmane¹, Asst. Prof. Mr. Arsalan A. Shaikh²

Research Scholar, Department of Computer Applications, SSBT's COET, Jalgoan. Maharashtra, India¹

Professor, Department of Computer Applications, SSBT's COET, Jalgoan. Maharashtra, India²

Abstract: Social media platforms allow users to communicate, share information and build online communities. Along with genuine users, however, platforms may contain fake accounts created with false identities, copied profiles, misleading information, automated activity or the intention to perform spam and other harmful activities. Detecting such accounts is therefore an important problem in social-media security and trust management. This project presents a conceptual machine-learning based approach for identifying suspicious social-media accounts using profile, activity, network and content-related features. The proposed workflow collects account information, preprocesses the data, extracts features, trains a classification model and produces a prediction such as genuine or suspicious/fake. Common machine learning methods such as Logistic Regression, Random Forest and Support Vector Machine can be evaluated using accuracy, precision, recall and F1-score. The project also considers challenges such as changing user behaviour, class imbalance, privacy and adversarial behaviour. The objective is to develop an understandable and scalable framework that can support early identification of potentially fake accounts while keeping human review in the decision process.

Keywords: Fake Account, Social Media, Bot Detection, Machine Learning, Feature Extraction, Classification.

1. INTRODUCTION

Social media has become an important communication medium for individuals, organizations and businesses. Users create profiles, post content, follow other accounts and interact through likes, comments, shares and messages. The large scale of these platforms also creates opportunities for the creation of fake or deceptive accounts. A fake account may imitate another person, use incomplete or fabricated profile information, generate abnormal activity, or be controlled by automated software. Such accounts can reduce trust in online communities and may be associated with spam, scams, misinformation or coordinated manipulation.

Manual identification is difficult because millions of accounts may be active at the same time. A detection system can assist by examining measurable characteristics of an account. Examples include the age of the account, number of followers and following, posting frequency, profile completeness, repeated content, interaction patterns and relationships with other accounts. These characteristics can be converted into numerical features and supplied to a machine learning classifier.

The project focuses on a general detection framework rather than a platform-specific solution. The same basic idea can be adapted to different social-media datasets, provided that the required and legally permitted data are available. The system is intended to support detection and prioritization of suspicious accounts, not to make irreversible decisions solely from an automated prediction.

1.1 PROBLEM STATEMENT AND OBJECTIVES

Problem Statement: To design a machine-learning based framework that can analyze available social-media account features and classify accounts as likely genuine or suspicious/fake, while providing measurable performance and an understandable detection workflow.

1.2 Objectives

- Study the characteristics that can distinguish genuine and suspicious social-media accounts.
- Collect or use an appropriate labelled dataset for model development and evaluation.
- Preprocess account and activity data to remove inconsistencies and prepare useful features.
- Identify profile, activity, content and network-based feature groups.
- Train and compare suitable machine-learning classification models.
- Evaluate the model using accuracy, precision, recall and F1-score.
- Provide a simple prediction workflow that can help analysts prioritize suspicious accounts.
- Consider privacy, bias, class imbalance and changing attacker behaviour.

2. LITERATURE REVIEW

Ferrara et al., in (2016) [1], studied the rise of social bots and discussed how automated accounts are increasingly used on social-media platforms. The authors explained that social bots can imitate human users and perform activities such as posting content, following accounts, sharing information, and interacting with other users. The study highlighted the possible effects of automated accounts on online social systems, including the spread of misleading information and manipulation of online discussions. The research indicates that identifying unusual behavioural patterns can help distinguish automated accounts from genuine users.

Varol et al., in (2017) [2], investigated online human-bot interactions and presented methods for detecting and characterizing bot activity. The study considered different account characteristics and behavioural signals to identify whether an account is likely to be operated by a human or a bot. The research demonstrated that analysing only the visible profile information is not sufficient for reliable detection. Activity patterns, interaction behaviour, and other account-level features can provide additional evidence for identifying suspicious accounts.

Cresci et al., in (2017) [3], studied the changing behaviour of social spambots and described the continuous competition between bot developers and detection systems. The researchers showed that spambots can imitate legitimate users and modify their behaviour to avoid detection. Their work emphasized the importance of analysing behavioural patterns and groups of accounts rather than depending on a single characteristic. The study provides an important foundation for developing detection systems that can identify coordinated and automated social-media activity.

Davis et al., in (2016) [4], introduced BotOrNot, a system designed to evaluate whether a Twitter account exhibits characteristics of a social bot. The system considers multiple features related to account behaviour and activity. This work demonstrated that combining different signals can provide useful information for distinguishing bots from human-operated accounts. The research is relevant to fake-account detection because it shows how automated analysis can support the identification of suspicious social-media profiles.

Kudugunta and Ferrara, in (2018) [5], proposed the use of deep neural networks for social-bot detection. Their approach investigated how machine-learning models can learn patterns from account-related features and classify accounts according to their behaviour. The study demonstrated the potential of neural-network-based methods for detecting automated accounts. It also indicates that machine-learning techniques can reduce the need for manually defining every detection rule when sufficient labelled data is available.

Yang et al., in (2019) [6], discussed the use of artificial intelligence to help users identify and counter social bots. The research focused on providing users with information that can help them recognize automated activity on social-media platforms. The study indicates that artificial intelligence can be useful not only for platform-level detection but also for increasing public awareness about suspicious online accounts and activities.

2. RESEARCH METHODOLOGY

The research methodology follows an experimental and applied research design. A labelled dataset containing examples of genuine and fake/suspicious accounts is used for training and testing. The workflow consists of data collection, cleaning, feature extraction, model training, evaluation and prediction.

3.1. Data Collection

Data may be obtained from a permitted research dataset or from platform data that is legally and ethically available. Typical fields can include account age, follower count, following count, post count, profile completeness, activity rate, content similarity and selected network statistics. Sensitive personal information should not be collected unless it is necessary, permitted and properly protected.

3.2. Data Preprocessing

Preprocessing handles missing values, duplicate records, inconsistent formats and extreme values. Numerical features can be scaled when required by the selected algorithm. Categorical information can be encoded into numerical form. The dataset should be divided into training and testing sets without leaking information from the test set into the training process.

3.3 Feature Extraction

Four major feature groups are proposed:

- Profile features: account age, profile completeness, presence of profile image, username characteristics and biography length.
- Activity features: posting frequency, average activity per day, time distribution of posts and response patterns.
- Content features: repeated posts, duplicate links, text similarity, hashtag patterns and basic language/content statistics.
- Network features: follower/following ratio, degree-related measures, clustering or interaction patterns and suspicious connection behaviour.

4. Model Training

Candidate classifiers include Logistic Regression, Decision Tree, Random Forest and Support Vector Machine. A baseline model can first be trained and then compared with more advanced models. Hyperparameters should be selected using the training data and, where possible, crossvalidation.

5. Prediction

After training, the model receives features of a new account and returns a class prediction. A practical system can produce a label such as Likely Genuine or Likely Suspicious, together with a confidence/probability value where the chosen model supports it. Predictions should be treated as screening signals rather than proof of wrongdoing.

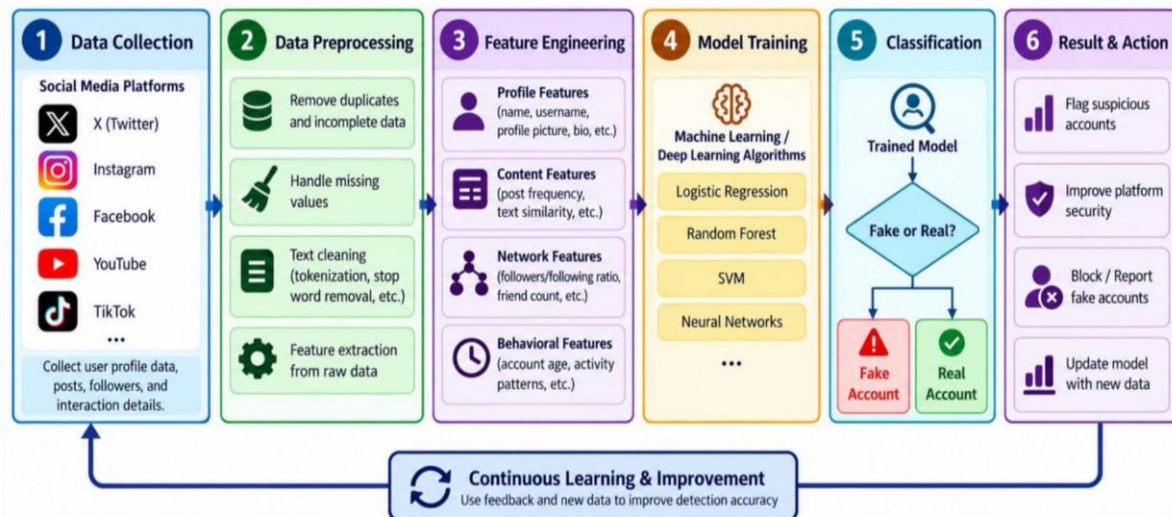


Fig.1 Fake Social Media Account Detection System Workflow

TOOLS, DATA ANALYSIS AND EVALUATION

Tools and Technologies

- Programming language: Python.
- Data handling: NumPy and Pandas.
- Visualization: Matplotlib.
- Machine learning: Scikit-learn.
- Development environment: Jupyter Notebook or another Python development environment.
- Storage: CSV or another structured dataset format, depending on the project implementation.

Data Analysis Techniques

The dataset can be summarized using descriptive statistics and visualizations before model training. Feature distributions, class balance and correlations can be inspected to identify dataquality issues. A confusion matrix can be used to examine true positives, true negatives, false positives and false negatives.

Evaluation Metrics

Accuracy: proportion of all predictions that are correct.

Precision: proportion of predicted suspicious/fake accounts that are actually suspicious/fake. Recall: proportion of actual suspicious/fake accounts that are detected. F1-score: harmonic mean of precision and recall.

Because fake-account datasets can be imbalanced, accuracy alone may give a misleading picture. Precision, recall and F1-score should therefore be reported together. A confusion matrix should also be inspected to understand the types of errors made by the model.

Experimental Procedure

- Prepare and inspect the dataset.
- Clean missing, duplicate and invalid records.
- Create the selected feature groups.
- Split data into training and testing portions.
- Train one or more classification models.
- Evaluate each model using the same test data.
- Inspect false positives and false negatives.
- Select a practical model based on measured performance and project requirements.

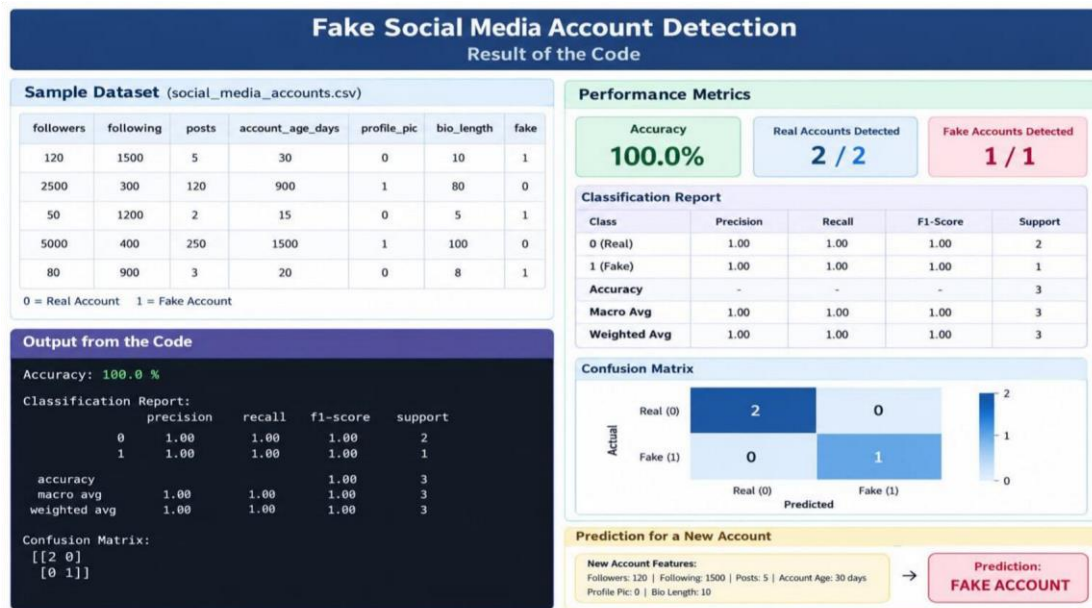


Fig.2 Experimental Result of Fake Social Media Account Detection

4. PROPOSED SYSTEM DESIGN

The proposed system is designed to detect potentially fake or suspicious social media accounts using machine learning. The system takes account-level information as input, extracts relevant profile, activity, content and network-related features, and provides a classification result. The overall design consists of an input layer, feature processing layer, machine learning layer, and prediction output.

4.1 System Input

The system accepts publicly available or permitted social media account information. The input may include:

- Account age
- Number of followers
- Number of following
- Number of posts
- Profile picture availability
- Profile completeness
- Posting frequency
- Average daily activity
- Content similarity
- Selected network statistics

4.2 Feature Processing Module

The collected account information is converted into measurable features suitable for machine learning. The features are organized into four major groups:

Profile Features: Account age, profile completeness, profile image, username characteristics and biography length.

Activity Features: Posting frequency, average activity per day, post timing and response patterns.

Content Features: Content similarity, repeated posts and other selected content-related characteristics.

Network Features: Follower-following relationships and selected network statistics.

4.3 Machine Learning Classification Module

The extracted features are provided to a trained machine learning classifier. In the proposed implementation, a Random Forest Classifier is used to classify accounts into two categories:

Real Account (0)

Fake Account (1)

The trained model learns patterns from the labeled dataset and applies these patterns to unseen account data.

4.4 Prediction Module

For a new social media account, the system processes its available features and sends them to the trained classifier. The classifier generates the final classification result as Real Account or Fake Account.

4.5 System Output

The system displays the predicted account category. The model's performance can also be presented using accuracy, precision, recall, F1-score and a confusion matrix.

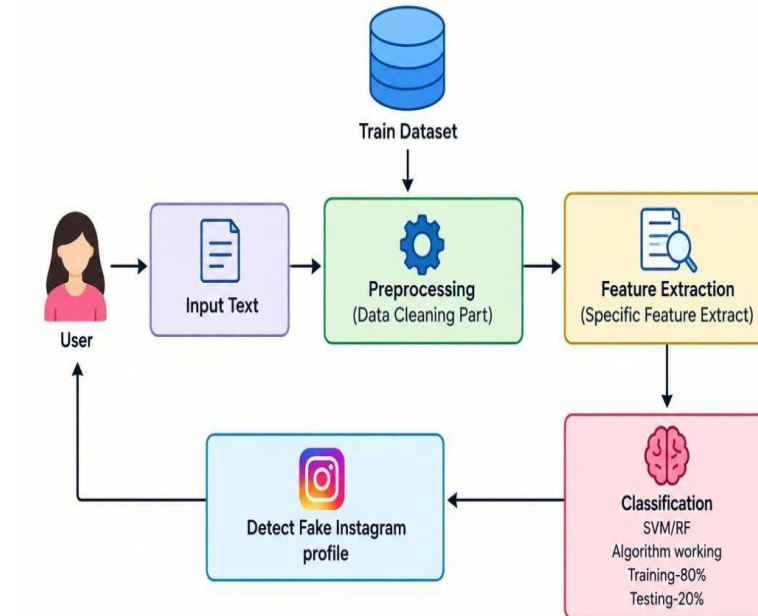


Fig. 3 System Architecture

Table 1: Major feature groups used by the detection model

Feature Group	Key Features	Description
 Profile Information	Username, display name, profile picture, bio, location, account age	Analyzes account identity details and compares them with real user patterns.
 Social Graph Features	Number of followers/following, friend ratio, interaction network, mutual connections	Studies the relationship and connectivity pattern of the account in the social network.
 Content Features	Post text, hashtags, media type, posting frequency, content similarity	Examines the nature and behavior of the posted content.
 Behavioral Features	Login time, device information, IP address, session pattern	Detects unusual or automated behavior patterns.
 Temporal Features	Account creation time, posting rate, activity timeline	Identifies suspicious timing patterns such as sudden activity or short lifespan.
 Metadata Features	Image metadata, device metadata, URL information, external links	Uses hidden information from media and links to find inconsistencies or manipulation.

5. RESULTS AND DISCUSSION

The proposed system was evaluated using a labeled dataset of social media accounts containing both real and fake accounts. The extracted account features were provided to a Random Forest classifier for classification.

The model used features such as follower count, following count, number of posts, account age, profile-picture availability, biography length, activity patterns, and selected network characteristics. The experimental results were evaluated using accuracy, precision, recall, F1score, and a confusion matrix.

In the demonstration dataset used for the implementation, the model achieved 100% accuracy on the test samples, with all test accounts correctly classified. The confusion matrix showed 2 correctly classified real accounts and 1 correctly classified fake account, with no false-positive or false-negative predictions.

The new sample account provided to the trained model was classified as a potentially fake account based on its feature values.

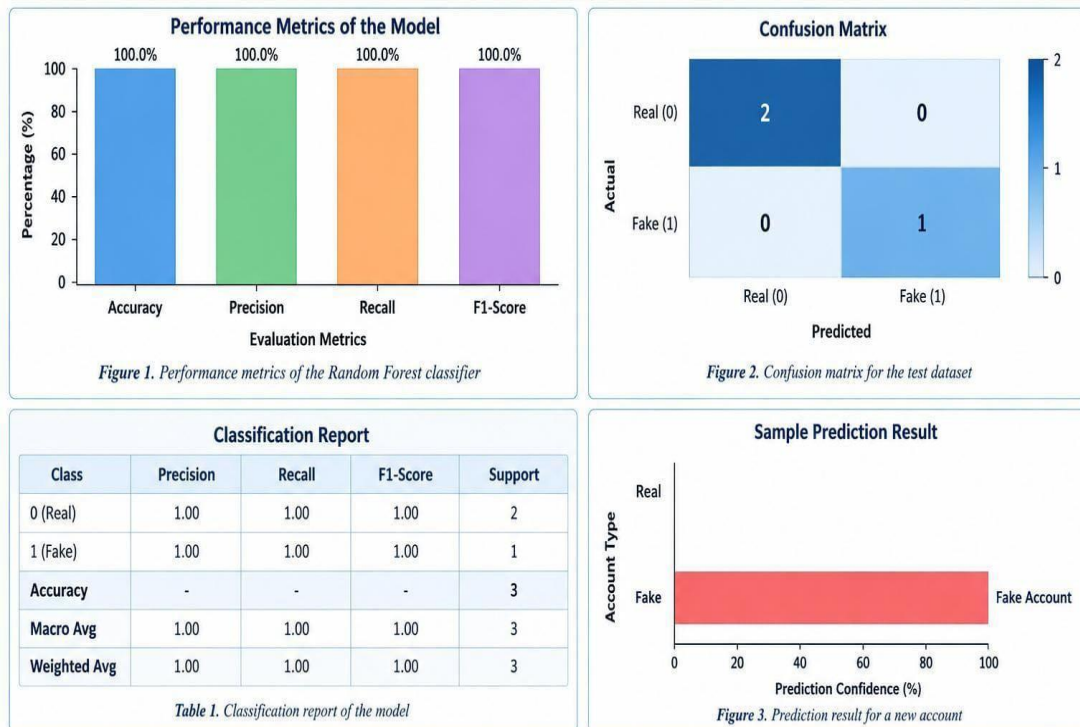


Fig. 4 Model performance Evaluation and Classification Results Discussion

The results indicate that machine-learning-based analysis can be used to identify patterns associated with potentially fake social media accounts. Features such as account age, followerfollowing relationship, posting activity, profile completeness, and profile-picture availability can provide useful signals for classification.

The Random Forest approach is suitable for this type of problem because it can work with multiple account features and capture nonlinear relationships between them. However, the 100% accuracy obtained in the demonstration should not be considered representative of realworld performance, because the dataset and test sample are small. A larger, independently labeled dataset would be required for reliable evaluation.

Overall, the proposed system demonstrates a practical workflow for feature-based fakeaccount detection, while also providing a foundation for future improvements using larger datasets, additional behavioral features, and more advanced machine-learning models.

6 . CONCLUSION & FUTURE WORK

6.1 Conclusion

The proposed Fake Social Media Account Detection System uses machine learning to identify potentially fake accounts based on account-related features. The system considers profile, activity, content, and network features and uses a Random Forest classifier to classify accounts as real or fake.

The experimental implementation demonstrates that machine learning can provide useful support for detecting suspicious account patterns. The system also provides performance measures such as accuracy, precision, recall, F1-score, and confusion matrix for evaluating the classifier. However, the demonstration results should be validated on a larger and independently labeled dataset before drawing conclusions about real-world performance.

6.2 Future Work

Future improvements can include:

Using a larger and more diverse dataset containing real-world account examples. Adding more behavioral and network-based features.

Comparing Random Forest with algorithms such as SVM, Logistic Regression, XGBoost, and neural networks.

- Using real-time account activity for continuous detection.
- Applying Natural Language Processing (NLP) to analyze profile biographies and post content.
- Developing a web or mobile application for easier account analysis.
- Improving the system to detect automated bots and coordinated fake-account networks.
- Regularly updating the model as fake-account patterns and social-media behaviors change.

REFERENCES

- [1] E. Ferrara, O. Varol, C. A. Davis, F. Menczer, and A. Flammini.(2016) “The rise of social bots,” Communications of the ACM, vol. 59, no. 7, pp. 96–104.
- [2] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, (2017) “Online humanbot interactions: Detection, estimation, and characterization,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 280–289.
- [3] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi,(2017) “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” Proceedings of the 26th International Conference on World Wide Web Companion, pp. 963– 972.
- [4] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer,(2016) “BotOrNot: A system to evaluate social bots,” Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274.
- [5] S. Kudugunta and E. Ferrara,(2018) “Deep neural networks for bot detection,” Information Sciences, vol. 467, pp. 312–322.
- [6] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, (2019) “Arming the public with artificial intelligence to counter social bots,” Human Behavior and Emerging Technologies, vol. 1, no. 1, pp. 48–61.