



Hybrid Feature-Based Comparative Analysis of Supervised and Unsupervised ML-Based Intelligent Bot Detection Model for Social Media Data

Siddhi Kale¹, Vishal Shirsath², Pradip Salve³ and Akriti Biswas⁴

^{1,2,3,4} Ajeenkya DY Patil University, Pune, India

siddhi.kale2503@gmail.com*
vishal.shirsath@adypu.edu.in
akritibiswas310@gmail.com
pradip.salve@adypu.edu.in

Abstract. Social media has become an important part of how people communicate and share information. However, the increasing use of it has led to rise in fake and automated bot accounts. These accounts create miscommunication and often spread fake news, spam, misleading information. Since bots are becoming difficult to identify, detecting them has become a necessary problem. In this study, we have developed a bot detection model - approach by combining text-based end user behaviour features. TF-IDF is used to analyse the text, along with features such as account information, engagement ratio, retweet activity, and mention patterns. 5 machine learning models were tested: Random Forest, Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and K-Means clustering. Among these models, Random Forest gave the best result, with an accuracy of 90%, while K-Means showed lower performance. The results show that using both text and user behavioural features can help improve bot detection and can support the development of safer and more reliable social media platforms which can be used by everyone.

Keywords: Bot Detection, Machine Learning, Supervised Learning, Unsupervised Learning, Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, K-Means, TF-IDF, Confusion Matrix, ROC-Curve, F1-Score

1 Introduction

Social media platforms have become an important part of everyday life. They allow people to communicate and share Information quickly with others around the world. However, the growing use of these platforms has also led to an increase in automated accounts commonly known as bots. These accounts are often used to spread spam, misinformation, miscommunication, manipulated content, which can affect the information which is shared online [2][5][15]. As bots may behave like real users, identifying them accurately has become more difficult machine learning methods can help with this problem by finding useful patterns in a large amount s of social media data. [2][17]. Previous studies have shown good results but many of them focus only on either text-based features or user behaviour features. Using only one type of feature

may not provide a complete picture of bot activity. So, to overcome and to address this issue, the proposed framework combines TF-IDF based text features with behavioural and numerical features search as follower count, engagement ratio, re-tweet activity, and mention patterns. The performance of the proposed approach is evaluated using 4 supervised learning algorithms: Random Forest, Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and along with unsupervised K-Means clustering approach.

1.1 Key Contributions of Proposed Solution:

1. Developed a hybrid feature- based framework by integrating textual, behavioral and numerical features to improve the detection of social media bot accounts.
2. Compared the performance of four supervised machine learning algorithms with an unsupervised K-means clustering model to evaluate their effectiveness.
3. Analyze the trends and limits of each model using multiple performance matrix to identify opportunities for further improvement.
4. Designed an efficient and interpretable workflow that can support reliable and scalable social media bot detection.

1.2 Problem Formulation

Let $D = \{x_1, x_2, \dots, x_n\}$ represent a social media dataset with each account containing textual and behavioral attributes. Textual features are transformed using TF-IDF, while behavioral features include follower log, engagement ratio, mention ratio. The objective of the proposed solution is to classify social media accounts into Bot or Human.

$$f(x_i) \rightarrow \{Bot, Human\} \quad (1)$$

Where, $f(x_i)$ represents the predicted class of a social media account, indicating whether an account belongs to the bot or human class.

2 Related Work

The increasing use of social media has encouraged extensive research on machine learning techniques for detecting automated bot accounts. Cresci et al. [1] demonstrated that behavioural and profile-based features are effective for identifying fake followers, achieving an accuracy of over 90%. Aljabri et al. [2] presented a comprehensive review of supervised, unsupervised, and deep learning approaches, concluding that supervised methods generally provide stronger classification performance. Arranz-Escudero et al. [3] and Lopez-Joya et al. [6] reported that combining textual, behavioural, and profile-based information improves the effectiveness of bot detection. Tanaka et al. [4] explored user interaction patterns for behavioural analysis, whereas Kulkarni [5] incorporated natural language processing with graph-based techniques to identify

coordinated bot activities. Baig [7] further demonstrated that hybrid machine learning approaches can achieve high detection accuracy. Based on these findings, integrating multiple feature types with supervised learning offers a practical and reliable solution for social media bot detection.

2.1 Comparison of Existing Bot Detection Approaches

Table 1. Comparison of Existing Social Media Bot Detection Methods

STUDY	METHOD USED	DATASE T USED	FEATUR E USED	REPOR TED ACCUR ACY	KEY ADVANTAGES
Cresci et al. [1]	Behavioral feature-based bot detection	Twitter fake follower dataset	Behavioral and profile-based features	>90% accuracy	Integrates textual and behavioural features for more effective bot detection.
Aljabri et al. [2]	Review of supervised, unsupervised, and deep learning approaches	Multiple benchmark social media datasets	Textual, behavioral, and deep learning features	80–95% accuracy range	Provides a comparative analysis of supervised and unsupervised machine learning models using hybrid features.
Arranz-Escudero et al. [3]	Multimodal machine learning framework	Social media benchmark datasets	Textual, behavioral, and profile-based features	Improved multimodal classification performance	Integrate multimodal features while maintaining a simple and interpret framework.
Tanaka et al. [4]	Web log behavioral analysis	Web log and user activity datasets	User-agent data, session activity, navigation behavior	Approximately 85–90% accuracy	Combines textual, behavioral, and engagement features to improve the detection.
Baig [7]	Hybrid adaptive transfer learning approach	Large-scale social media datasets	Hybrid supervised and unsupervised features	97% accuracy	Delivers strong performance without relying on complex, adaptive architectures.
DeeProBot [18]	Hybrid deep learning model	Twitter bot datasets	Deep neural textual and behavioral representations	High detection performance	Offers feature level inter ability without depending on deep learning models.
BotRGCN [19]	Graph neural network-based bot detection	Twitter graph datasets	Relational graph and network	Improved robustness in coordinate	Eliminates the need for graph-based methods by integrating textual

interaction
features

d bot
detection

and behavioral
features.

Comparative study highlights the modern bot detection approaches in increasingly adopt hybrid, deep learning and graph-based methods to enhance detection accuracy. Despite their strong performance, these approaches required substantial computational resources in large data. The proposed hybrid framework offers a computationally efficient alternative by integrating textual and behavioral features for maintaining reliable classification performance

3 Research Methodology

3.1 Dataset Overview

The experiments were performed using the publicly available Twitter Bot Detection Dataset available on Kaggle. The dataset contains Twitter accounts which are labelled and divided in two classes – bot and human, along with textual and behavioral features. The dataset contains 50000 records in which bot accounts are 25018 and human accounts are 24982. Since datasets are balanced, no oversampling methods were required.

Dataset: Twitter Bot Detection Dataset

Dataset Link: <https://www.kaggle.com/datasets/goyaladi/twitter-bot-detection-dataset>

3.2 Data Cleaning and Preprocessing

We used various pre-processing methods on the dataset before training the model. All the text was converted to lowercase during pre-processing and empty text fields were replaced with blank strings. Numerical data was converted into an appropriate format for analysis and standardization. Any invalid numerical data, such as negative numbers in Mention_Count, and missing values were replaced with 0. We used stratified sampling for even class distribution throughout the model training.

3.3 Feature Engineering

In this process, meaningful features and relations between the data were extracted in order to build an efficient model.

Table 2. Hybrid Textual and Behavioral Features Used for Bot Detection

FEATURE	DESCRIPTION
Follower_Log	Log-transformed follower count used to reduce skewness in follower distribution
Retweet_Count	Number of retweets associated with user activity and engagement
Mention_Count	Frequency of mentions observed in user interactions.
Engagement_Ratio	Ratio derived from retweet activity and follower count to measure engagement intensity

Mention_Ratio	Mention activity normalized with respect to follower count
Verified	Verification status of the social media account
TF-IDF Features	Numerical textual representation generated using TF-IDF vectorization for tweet content analysis

3.4 Numerical Features

Numerical features were extracted to capture user interaction patterns. Logarithmic transformation was applied to follower count to reduce skewness, while engagement ratio and mention ratio were used to measure interaction and communication patterns. Verification status, retweet count, and mention count were also included as numerical features. The numerical features used in the proposed framework are mathematically represented as follows:

1. Follower_Log

$$Follower_Log = \log(1 + Follower.Count) \quad (2)$$

2. Retweet_Count

$$Retweet_Count = Retweet.Count \quad (3)$$

3. Mention_Count

$$Mention_Count = Mention.Count \quad (4)$$

4. Engagement_Ratio

$$Engagement_Ratio = \frac{Retweet.Count + 1}{Follower.Count + 1} \quad (5)$$

5. Mention_Ratio

$$Mention_Ratio = \frac{Mention.Count}{Follower.Count + 1} \quad (6)$$

6. Verified

$$\begin{aligned} Verified \\ = 1, \text{ if the account is verified } 0, \text{ if the account is not verified} \end{aligned} \quad (7)$$

3.5 Textual Features (TF-IDF)

Textual features were determined using the TF-IDF (Term Frequency–Inverse Document Frequency) for clearly highlighting frequently used terminologies by Twitter bot accounts [11][12]. During this process, the tweets were tokenized and further converted to numeric values where each terminology was assigned a specific weight based on its importance and relevance.

$$TF\text{-}IDF(t,d) = TF(t,d) \cdot \log\left(\frac{N}{DF(t)}\right) \quad (8)$$

3.6 Feature Integration

After determining respective features, they were integrated together for better analysis as real-world accounts need to be judged on multiple parameters before determining whether they are bot or human. This led to improvement in the model performance in real world scenarios and achieve greater accuracy.

$$[X = x_1, x_2, x_3, \dots, x_n] \quad (9)$$

3.7 Model Selection

Four supervised machine learning models (Random Forest, Support Vector Machine, Logistic Regression and Naive Bayes) and one unsupervised model (K-means clustering) were used. SVM handles high dimensional data effectively and is useful of binary classification tasks [9][10]. Logistic regression offers a probabilistic method for classification [10][11]. Similarly, Naive Bayes is highly effective for textual classification [11][12]. K-means clustering was used as a baseline method to evaluate performance without any labelled data. The classification task can be mathematically represented as:

$$\hat{y} = f(X) \quad (10)$$

3.8 Model Configuration

The dataset was split into 80:20 for training and testing, with random seed=123. Random Forest was implemented using 300 trees, mtry was set to the square root of total features and minimum node size = 5. SVM was implemented using an RBF Kernel. Logistic Regression and Naive Bayes were implemented using the standard configuration. K-means clustering implemented with K = 2 (bot and human classes).

3.9 Evaluation Metrics

Following are the evaluation metrics used to analyze model performance:

- **Accuracy:** indicates how correct the model works overall

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

- **Precision:** shows that out of these predicted bots how many are actual bots.

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

- **Recall:** tells us out of total bot accounts present how many were actually caught and identified by a model.

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

- **F1-score:** Following are the evaluation metrics used to analyze model performance.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (14)$$

- **Specificity:** highlights that out of total human accounts how many did the model correctly label as human.

$$Specificity = \frac{TN}{TN + FP}$$

(15)

- **ROC-AUC:** specifies the model’s ability to clearly differentiate between both the classes across different thresholds

$$ROC - AUC = \int_0^1 TPR(FPR)d(FPR)$$

(16)

3.10 Experimental Setup

The experiments were performed in the RStudio Environment using the R programming language on a system with a 13th Gen Intel(R) Core (TM) i5-1334U (1.30 GHz) processor, 16.0 GB RAM, and Windows 11 operating system. Libraries included are text2vec, ggplot2, caret, e1071, pROC, ranger and dplyr.

4 Experimental Evaluation

This section presents a detailed comparative analysis of all the five machine learning models used in the experiments for social media bot detection.

4.1 Comparative Performance Evaluation

After reviewing performance results of all 5 models, key findings included that supervised machine learning models performed considerably better than the unsupervised learning model. Random forest was the best performing model with an accuracy of 90 percent and 0.9685 AUC. The performance of different machine learning models was evaluated using following metrics summarized in the table below.

Table 3. Performance Comparison of Supervised Learning Models

Model	AUC	Accuracy	Precision	Recall	F1 Score	Specificity
Naïve Bayes	0.8726	0.817	0.7462	0.9606	0.8400	0.6738
SVM	0.8769	0.803	0.7601	0.8859	0.8180	0.7208
Logistic Regression	0.8764	0.817	0.7356	0.9894	0.8430	0.6448
Random Forest	0.9685	0.900	0.8547	0.9642	0.9061	0.8363

K-Means	0.500 1	0.5004	0.7000	0.00028	0.00056	0.99988
----------------	------------	--------	--------	---------	---------	---------

Shown below is the visual representation of ROC curves of all models:

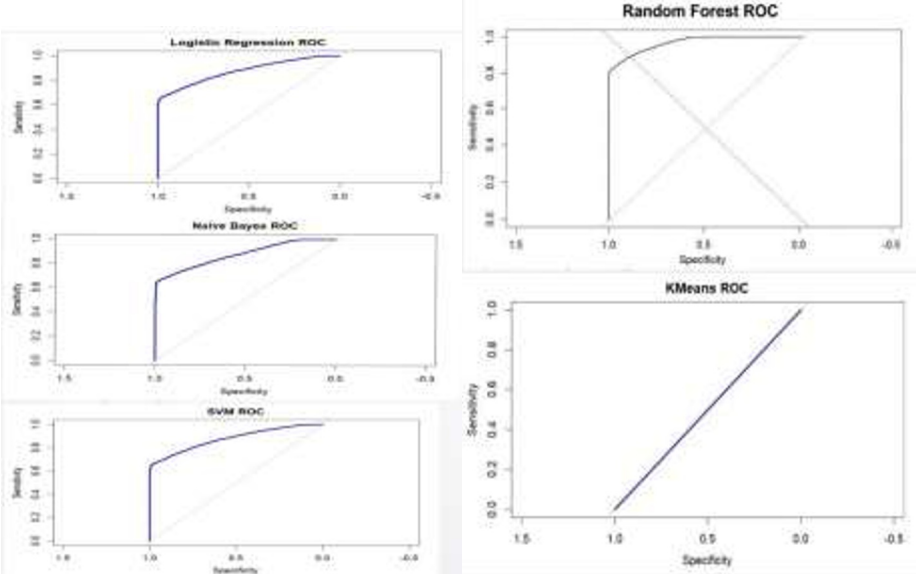


Fig.1. ROC curve of the 5 models

4.2 Explainable AI

To gain a better understanding of how the processed model reached its predictions, feature importance was examined using the Random Forest algorithm. The analysis showed that the behavioural features such as Mentioned Count, Mentioned Ratio, Follower Log, Retweet Count, had a greater influence on the classification results than many of the textual features. These attributes were particularly effective in capturing unusual interaction patterns and repetitive activities that are commonly associated with the bot accounts. The results also highlight the complementary role of both feature types. While behavioural features were more influential in distinguishing bot accounts from genuine users, TF-IDF based textual features helped capture meaningful patterns in the content of user posts. Combining these two sources of information not only improve the overall classification performance but also made the models decision easier to understand and interpret.

5 Conclusion & Future Scope

The results of this study show that combining text features such as TF-IDF with user behavioural features is useful for detecting social media bot accounts. The supervised learning models performed better than the unsupervised approach. Among all the models, Random Forest achieved the highest accuracy. The study also shows that using

different types of features together can improve bot detection. In the future this framework can be tested on larger and bigger datasets from some social media platforms such as LinkedIn, Instagram, and Facebook. Deep learning models can also be used to improve the performance of the system. Explainable AI (XAI) techniques such as SHAP and LIME can be added to help understand how the model makes its decisions and makes the results easier to explain.

5.1 Ethical And Responsible AI Considerations

Machine learning models used for detecting social media bots should be developed and used carefully because wrong predictions can affect real world users. Therefore, fairness, transparency and user privacy are important while building and using such systems. This proposed model framework uses text and user behavioural features instead of sensitive personal information, which helps reduce privacy concerns. Human review should be considered and should also be done before taking any serious action against an account marked as suspicious. In the future, Explainable AI techniques can be used to make the model's decision easier to understand. This can help make the bot detection systems more fair, transparent and reliable.

6 Acknowledgement

The authors of this paper would like to extend their sincere thanks to the Computer Science Engineering department of Ajeenkya D.Y. Patil University, Pune, for providing the necessary guidance and facilities required to carry out this research. The authors also thank their project guide and faculty members for their unwavering support and motivation throughout this research.

7 Disclosure of Interest

The authors declared that there are no conflicts of interest related to this research or its publication. They also confirm that they have no financial or personal relationships that could have influenced the work.

References

- [1] 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.
- [2] M. Aljabri, A. Alotaibi, and A. Alghamdi, "Machine learning-based social media bot detection: A review," *IEEE Access*, vol. 11, pp. 12345–12360, 2023.
- [3] G. Arranz-Escudero, et al. "Multimodal detection of social bots using machine learning techniques," *IEEE Access*, 2025.

- [4] T. Tanaka, H. Niibori, S. Li, S. Nomura, H. Kawashima, and K. Tsuda, "Bot detection model using user agent and user behavior for web log analysis," *Procedia Computer Science*, vol. 176, pp. 1621–1625, 2020.
- [5] W. J. Kulkarni, "Twitter bot detection using NLP and graph classification," Master's Project, San Jose State University, 2023.
- [6] S. Lopez-Joya, J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, "Exploring social bots: A feature-based approach to improve bot detection in social networks," *arXiv preprint arXiv:2411.06626*, 2024.
- [7] M. D. Baig, "Bot detection using a machine learning adaptive transfer approach," *Emerging Learning Technologies*, vol. 1, no. 1, pp. 20–29, 2024.
- [8] Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
- [9] Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* 20(3), 273–297 (1995)
- [10] Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1–22 (2010)
- [11] McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: *AAAI Workshop on Learning for Text Categorization* (1998)
- [12] Ramos, J.: Using TF-IDF to determine word relevance in document queries. In: *Proceedings of the First Instructional Conference on Machine Learning* (2003)
- [13] Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: BotOrNot : A system to evaluate social bots. In: *WWW Companion* (2016)
- [14] Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: Detection, estimation, and characterization. In: *ICWSM* (2017)
- [15] Ferrara, E., Varol, O., Davis, C.A., Menczer, F., Flammini, A.: The rise of social bots. *Communications of the ACM* 59(7), 96–104 (2016)
- [16] Gilani, Z., Kochmar, E., Crowcroft, J.: Classification of Twitter accounts into automated agents and human users. In: *ASONAM* (2017)
- [17] Kudugunta, S., Ferrara, E.: Deep neural networks for bot detection. *Information Sciences* 467, 312–322 (2018)
- [18] Hayawi, K., Shahriar, S., Serhani, M.A., Alashaikh, A., Mathew, S.: DeeProBot: A hybrid deep learning model for social bot detection. *Expert Systems with Applications* 187, 115945 (2022)
- [19] Feng, S., Tan, Z., et al.: BotRGCN: Twitter bot detection with relational graph convolutional networks. In: *ASONAM* (2021)
- [20] Mbona, I.R., Eloff, J.H.P.: Feature selection using Benford's law for bot detection. *Computers & Security* 101, 102118 (2021)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

