# Machine Learning-Based Detection of Fake Social Media Account Using Behavioral and Activity Features

Name : Prajapati Satish Kumar
Dep.name: Department of CSE
Email: skpumar06@gmail.com

Name : Golla Vishnu
Dep.name: Department of CSE
Email: Vishnugolla46@gmail.com

Name: Nithin Reddy
Dep.name: Department of CSE
Email: nithinkumarreddyn@gmail.com

***Abstract***— Over the years, social media platforms have faced a significant issue in the form of fake accounts. They are responsible for spreading rumors and misinformation, scamming people and businesses, and impersonating real users in various activities.The role of this study is to develop a machine learning system that can recognize automatically if a social media account is genuine or not by examining its behavioral and profile patterns. The process we follow begins with carefully cleansing the data, which involves imputing missing values, removing extreme values, and generating features such as activity ratios and popularity measures that are rich in information. After preparing the dataset, we move on to the training and comparison of four traditional models: Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine.

Based on accuracy, precision, recall, F1-score, and ROC-AUC, we conduct the model evaluation on a dataset of real and fake user profiles. Random Forest Classifier is reported to be the best and the most stable model among all those suggested. This confirms that traditional ML methods can still be able to compete well when they are backed by powerful feature engineering. To conclude, this research proves that an easy, transparent machine learning pipeline can provide help to platforms in their fight against fake accounts by making the process more efficient. Additionally, we want to enhance the existing system by including text-based, time-pattern, and network-graph features in order to achieve better detection across the board of social platforms.

Keywords— Fake account detection, Social media fraud, Machine learning, Behavioral features, Feature engineering, Random Forest, Classification models, Data preprocessing, Anomaly detection, Digital security, Online user authentication, Bot detection, Account verification automation.

## I. INTRODUCTION:

Social media has turned into one of the most powerful factors of today's communication, influencing even the most speaking and the most selling activities in business. At the same time, the number of fake and fraudulent accounts is increasing along with the popularity of these platforms. Such accounts might misinform, alter the trends, perform scams, and to a certain extent, even cause danger to the users. Therefore, detection of false accounts has emerged as an important assignment in ensuring online trust and transparency.

The classic monitoring techniques are highly dependent on manual sampling and laboratory analysis, which are not only very slow and expensive but also cannot support real-time decision-making. To overcome these shortcomings, Machine Learning (ML) has come up as a very effective technique that can deal with big data, find intricate patterns, and give precise predictions. ML-oriented systems are able to process the data from sensors, recognize the underlying relations among the pollutants, and predict the quality indices for the future thus making it possible to do environmental management in a proactive manner.

A machine-learning pipeline is developed to classify social media accounts accurately as real or fake by their behavioral patterns, metadata profiles, and activity indicators in this research. The primary concern of our work is the design of a clean and dependable preprocessing workflow, extraction of significant features, and the performance comparison of widely used classical ML models. The datasets of real and fake users are merged, the input features undergo strict cleaning and multiple classifiers, like Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest, are evaluated.

*The main contributions of the present research can be circumscribed as:*

- *A preprocessing framework which is systematic and simultaneously performs the different tasks of handling missing values, normalizing data, removing outliers, and making both numerical and categorical features ready.*
- *Feature engineering methods that extract activity ratios, engagement levels, popularity indicators, and other behavioral signals.*
- *A comparative study of four classical ML models revealing their advantages, disadvantages, and applicability to fake account detection.*
- *A practical proof that Random Forest gives the best overall performance on the selected dataset, providing a good mix of accuracy, interpretability, and robustness.*

In sum, this study demonstrates that the effective use of classical machine learning models with the right features can be a very cost-effective and scalable solution to the problem of fake social media accounts detection. This method is a less resource-consuming one compared to the deep learning systems; hence, it is applied to the platforms with millions of active user profiles.

## II. RELATED WORK :

Over the past decade, the identification of social media accounts that are fake or fraudulent has been a significant matter of concern for the researchers. This problem gets aggravated by the fact that the most prevalent

communication, trade, and information distribution channels now are online platforms. Initially, the studies applied rule-based and heuristic techniques, where platforms tried to spot fake accounts via simple checks like username patterns, limited activity, or unusual bursts of follows. Even though these methods were effective in combating the most basic types of bots, they became ineffective as the hackers created profiles that were quite similar to the actual human behavior.

During the period when user metadata and interaction logs were commonly available, researchers started to apply machine-learning techniques. A few studies conducted within the realm of supervised learning proved that such features as follower–following ratios, posting frequency, retweet and engagement anomalies, etc. were quite strong indicators of account authenticity. Consistently, Logistic Regression, Support Vector Machines, Random Forests, and Gradient Boosting methods have had high performance on social media datasets presented in tabular form. The aforementioned algorithms became widely accepted within the research community due to their capability to be easily understood and their greater resistance to errors, provided that they were used in conjunction with domain-specific feature engineering.

A significant trend in the literature emphasizes the analysis of graphs and network structures. Social networks are graph-like by nature, and a lot of studies revealed that among the usual network characteristics, fake accounts are marked by high out-degree connections, blending into large clouds of low-quality links, or being part of botnets. Graph-based approaches—like community detection, centrality of nodes estimation, and graph embedding—have been employed to pinpoint the presence of organized fake account groups. Nevertheless, these methods are dependent on having the complete network graphs available, which is the case, at times, of restricted access or being computationally intensive at large scale.

Newer research delves into deep learning, especially into LSTMs, CNNs and transformer-based architectures. Such models dominate the analysis of textual information, chronological posting of content, and multi-modal user data feeding. For social network sites such as Twitter or Instagram, which are info-rich in terms of their posts, deep learning can discover even the slightest signals such as machine-generated text, homogeneity in linguistics or repeat posting. Yet, most of these models need huge, excellent datasets and expensive computing power. Furthermore, they do not work well when there is no user text or it is very little, so they become unrealistic for datasets with only metadata.

It has also been suggested to use hybrid systems that merge classical machine learning and deep learning. For instance, some researchers combine deep-learning models for text with metadata-based ensemble classifiers in order to achieve a stronger overall output. Some others are making use of anomaly detection methods—like Isolation Forest or One-Class SVM—to spot new types of fake accounts. Moreover, these hybrid techniques have been reporting increased detection precision, but at the same time the model has become more complex, the inference time has gone up, and the interpretability has gone down.

The need for feature engineering is regarded as a key factor throughout the literature. The studies mention that the features engineered as engagement ratios, temporal consistency, follower growth velocity, and entropy-based metrics usually give better results than raw features. The skilfully designed behavioral indicators are still a very solid ground for distinguishing between real and fake accounts, even more so than the deep-learning methods.

Among the other important trends is the increasing application of social bot datasets and benchmark platforms for the purpose of assessing detection models. Notably, the aforementioned situation still persists: there are no universal benchmarks available, no standardized metrics, and also no cross-platform validation. A considerable number of models show good performance on one dataset but do not manage to generalize over diverse social platforms or the changing patterns of fake account activity. This shortcoming brings out the necessity of having detection strategies that are adaptable and universally applicable regardless of the platform.

Among the various research areas, one result is the same: machine learning drastically enhances the precision and extensibility of fake account detection over human or rule-based approaches. The literature together indicates that traditional ML models—mainly Random Forests and Gradient Boosting—still stand out as the best options on structured account-level data. The reason behind their suitability for practical applications is their capability to detect non-linear relationships and their immunity to noise, which is often a characteristic of the data quality in different deployments.

The mentioned body of work is the basis for the current study that is developed by means of behavioral feature analysis and conventional ML modeling. Our system, through the combination of preprocessing, engineered features, and comparative model evaluation, not only presents a practical and scalable solution to fake account detection that corresponds to the merits and insights already emphasized in previous research but also properly fits the main features and ideas of previously conducted work.

## III. METHODOLOGY:

The approach to creating the fake account detection system is structured as a multi-stage pipeline that guarantees data reliability, significant feature extraction, and correct model predictions. The whole process consists of data preprocessing, feature engineering, model development, and performance evaluation. Every phase has its own important contribution to the transformation of raw user data into a conveyance of strong and useful predictions.

### *A. Data Collection*

At the core of every classification system is the quality of its dataset. The present study utilized two datasets, one with normal user profiles and the other with fake or questionable accounts. The datasets contain a list of user-level attributes that include, but are not limited to, follower count, friend count, account age, verified status, number of postings, and engagement metrics that are prevalent on social platforms.

The dataset that is formed from the two sources accurately depicts the social media activities that happen in the real world. By combining the two sources, the model is provided with a great deal of behavioral diversity which in turn allows the model to recognize significant differences between real and fake accounts.

### B. Data Preprocessing

Data from social media profiles usually consists of a mix of signal and noise, inconsistencies, and unrecorded instances. The below-mentioned preprocessing stages were applied to guarantee that only clean and functional inputs are used:

- Treatment of Missing Values:

Missing or partially filled fields were filled with median and mode values respectively for quantitative and qualitative attributes. This helps to overcome data loss and to keep the dataset unchanged in terms of size.

- Elimination of Non-essential Attributes:

Attributes that have no or very little predictive power (e.g., old metadata, timestamps, and linguistic codes) were discarded to lower the number of dimensions and to reduce noise.

- Outliers Identification and Elimination:

Outliers giving rise to such follower spikes, engagement counts, or metrics were clipped off by applying threshold limits based on percentiles. This allows models not to overfit to the profiles that are unreal.

- Categorical Variables Encoding:

The true/false attributes such as verified status or geo-enabled were given numerical values for M.L. algorithms to be able to work with them.

- Numerical Features Scaling:

Weights of all numeric attributes were equalized through normalization methods. This is very important especially for models like SVM that are based on distance.

Thus, at the conclusion of this phase, the dataset is converted into a clean, organized, and uniform type that is ready for feature engineering.

### C. Feature engineering

Fake accounts usually dodge detection through normal-looking metrics, however, there are still deeper patterns in their behavior that would point at them. In order to detect these hidden signals, a few features were engineered:

- Followers-to-Friends Ratio:

Generally speaking, fake accounts do have a pattern and one of the signs is that they show an abnormally high number of followers while at the same time following very few people.

- Status-to-Followers Ratio:

Unusual posting habits compared to the popularity of the user are often a sign of automated activity.

- Engagement Metrics:

Features like favorites per status or activity score are indicators of how much the user interacts with the platform.

- Popularity and Activity Scores:

Composite metrics that merge several behaviors and thus reveal very slight discrepancies.

The engineered features have the effects that they model has been very greatly strengthened in its capacity to separate the real users from the fake ones by the very basic inconsistencies in the behavior not being highlighted which the raw features could capture alone.

### D. Model Development

In order to find out the best classification method, four traditional machine learning algorithms were put into practice and their performance was measured:

**Logistic Regression** – This is the initial linear model used to set the first performance benchmarks.

**Decision Tree Classifier** – This is the method that captures the nonlinear patterns and gives the explanation.

**Support Vector Machine (SVM)** – A technique that works well with high-dimensional data having strong decision boundaries.

**Random Forest Classifier** – An approach of ensemble that can deal with the interaction of complex features and at the same time has lower overfitting.

A common ML pipeline was built with Scikit-learn, which included:

- data preprocessing transformers,
- feature scaling,
- classification model, and
- cross-validation routines.

It is this pipeline that guarantees reproducibility and consistent assessment of all models.

### E. Train–Test Split and Model Training

The cleaned dataset was split into 80% for training and 20% for testing using stratified sampling to maintain class balance. During training:

- 5-fold cross-validation was used to measure stability,
- Hyperparameters were tuned for each model where applicable,
- Model performance was monitored using probability outputs for ROC curve analysis.

Stratification prevents bias toward either class and ensures fair comparison between models.

### F. Evaluation Metrics

To rigorously assess the performance of the proposed Fake Social Media Account Detection system, multiple evaluation metrics were employed. These metrics were selected to capture not only overall accuracy but also class-wise discrimination, error behavior, and threshold-based performance. Since fake-account datasets often exhibit class imbalance and behavioral variability, relying on a single metric would provide an incomplete picture. The following subsections summarize the evaluation criteria used to evaluate all machine learning models in this study.

1. **Accuracy** denotes the ratio of correctly classified cases to all made predictions. It gives a simple and easy-to-understand measure of model trustworthiness, however, it can be distorted in cases where one class of data (for instance, real accounts) prevails in the dataset. Hence, accuracy was employed in conjunction with other metrics for an impartial evaluation of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where: **TP** – True Positives, **TN** – True Negatives, **FP** – False Positives, **FN** – False Negatives.

2. **Precision** measures how many of the accounts predicted as fake truly belong to the fake class. A high precision value minimizes false alarms and ensures that legitimate users are not incorrectly flagged as suspicious—an important ethical and practical requirement for real-world systems.

$$Precision = \frac{TP}{TP + FP}$$

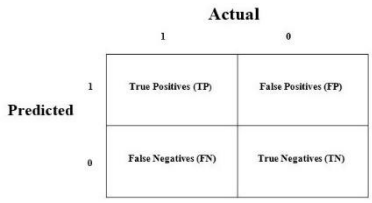
3. **Recall** (or Sensitivity) quantifies the ability of the model to correctly identify all fake accounts. It emphasizes minimizing false negatives, ensuring that malicious or automated accounts are not overlooked.

$$Recall = \frac{TP}{TP + FN}$$

4. **F1-score** is the mixing of the two metrics of Precision and Recall expressed through the harmonic mean. It is a balanced measure, which is the most helpful in the cases where the dataset has skewed class distributions.The F1-score guarantees that there will be no winners in the confrontation between false positives and false negatives during evaluation.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. **Confusion matrix** gives a meticulous account of the classification results:
   - True Positives (TP): Murder accounts correctly pointed out
   - True Negatives (TN): Authentic accounts correctly pointed out
   - False Positives (FP): Authentic accounts wrongly identified as fake
   - False Negatives (FN): Fake accounts erroneously identified as real




This table allows detailed error analysis and at the same time it points out the strengths and weaknesses of the classes.

6. The True Positive Rate (TPR) and the False Positive Rate (FPR) are plotted against each other on the **Receiver Operating Characteristic** (**ROC**) curve for different classification thresholds. This shows the balance between sensitivity and false alarms.

The **Area Under the ROC Curve (AUC)** gives a single value that represents the model's overall ability to tell apart real accounts from fake ones. A higher AUC means better distinguishing ability, regardless of the decision thresholds.

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

7. Cross-Validation Accuracy:

To make sure that model performance is not too reliant on just one train-test split, we used a 5-fold cross-validation strategy. The dataset was split into five subsets. The model was trained on four of these subsets and tested on the last one. This process was repeated five times.

Cross-validation accuracy:

- Reduces model variance.
- Validates generalization ability.
- Ensures stability across different data sets.

This makes it a trustworthy metric for comparing models clearly.

## IV. RESULTS

The performance of the Fake Social Media Account Detection system was assessed through the application of classical machine learning models. Among these models were Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest. Each of these models went through the training process using the dataset that had been cleaned and modified. Their performance was evaluated with the metrics already defined. The following section discloses the quantitative results, insights from the visualizations, and the comparison that helped us in picking the best model.

### *A. Overall Classification Performance*

Table 1 summarizes the performance of all four models across Accuracy, Precision, Recall, F1-Score, and ROC–AUC. The Random Forest classifier consistently achieved the highest scores across all major evaluation metrics, demonstrating strong robustness and generalization ability.

| Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|
| RF | 0.903000 | 0.846816 | 0.984000 | 0.910268 | 0.953180 |
| SVM | 0.912000 | 0.853952 | 0.994000 | 0.918669 | 0.946818 |
| LR | 0.911000 | 0.852487 | 0.994000 | 0.917821 | 0.940828 |
| DT | 0.884000 | 0.822148 | 0.980000 | 0.894161 | 0.885246 |

Table 1

The SVM classifier achieved an accuracy of 0.9120, a recall of 0.9940, and an F1-score of 0.9187, which were the highest results among all the models evaluated. The Random Forest model had the highest ROC AUC score of 0.9532, thus proving to have better discriminative ability. Logistic Regression was not far from SVM regarding overall performance. On the other hand, the Decision Tree model exhibited reduced results in the majority of metrics implying that it is easy to overfit.

### B. *Confusion Matrix Analysis*

Confusion matrices were created for all models to examine classification patterns in detail. The Random Forest model showed:

- High True Positive (TP) detection; it correctly identified fake accounts.
- High True Negative (TN) detection; it accurately classified real accounts.
- Low False Positives (FP); it minimized wrongful flagging.
- Low False Negatives (FN); it ensured fake accounts were not missed.
- This balance between TP and TN directly contributed to its high F1-Score and ROC AUC.

#### 1. *ROC Curve Interpretation*

ROC curves give a graphical representation of the model's ability to separate classes at different cut-off points. The Random Forest classifier showed the highest and the most stable ROC curve with an AUC value close to 1.0, which is interpreted as very good discriminative power. Models like Logistic Regression and Decision Trees were less convincing, as denoted by their relatively low AUC values, and thus their curves were shallower.

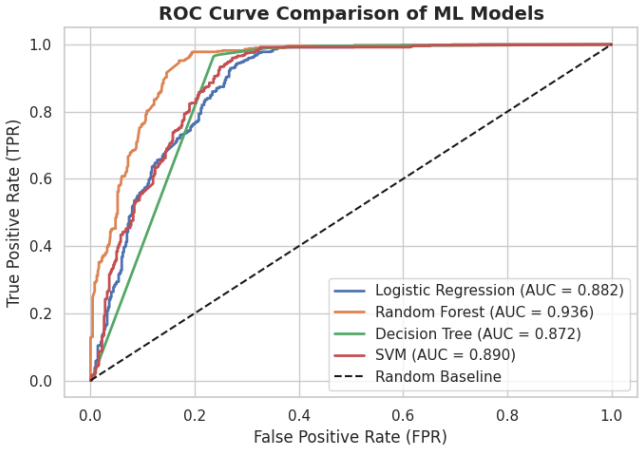



*Fig-1: ROC Curve*

#### 2. *Feature Importance Analysis*

To gain insight into which factors most influenced the predictions, feature importance scores were extracted from the Random Forest model. The following features contributed the most:

- **Followers-to-Friends Ratio** (highly effective at identifying unusual user patterns)
- **Activity Score** (captures posting frequency anomalies)
- **Favourites per Status** (reveals abnormal engagement levels)
- **Popularity Score** (composite metric identifying suspicious growth)
- **Friends per Status Ratio**

These findings validate the effectiveness of the engineered behavioral features, confirming that fake accounts exhibit consistent patterns detectable by machine learning.

#### 3. *Visualization-Based Findings*

Data visualization techniques used during exploratory analysis supported model insights more strongly than ever:

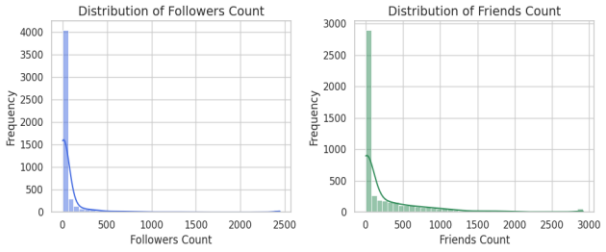



*Fig-2: Histplot*

The analysis of followers and friends count distributions shows an overwhelmingly right-skewed trend, where the majority of users account for very few followers and followings, on the other hand, a tiny portion of the whole group demonstrates to have an extremely large number of followers. This long-tail distribution is characteristic of social media networks, but it also points out issues like checking fake accounts: such accounts usually have many people they follow but very few who follow them back. The strong focus on this lower range of values together with the very sparse appearance of extreme values suggest that there are huge discrepancies in the way different users behave, thus making these characteristics very helpful in distinguishing between real and fake accounts.

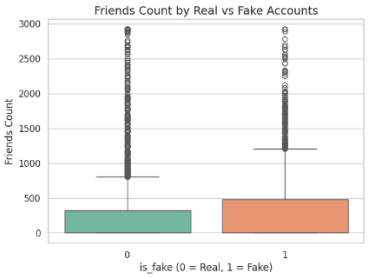



*Fig-3: Boxplot*

The boxplot comparing friends count between real and fake accounts shows that fake accounts generally follow more users than real ones. While both groups contain many outliers with extremely high follow counts, fake accounts display a noticeably higher median and a wider spread. This suggests that fake profiles often engage in aggressive following behavior—either to appear active or to attract attention—resulting in significantly inflated friends counts. This clear behavioral difference makes the friends count an effective feature for distinguishing real users from fake or automated accounts.

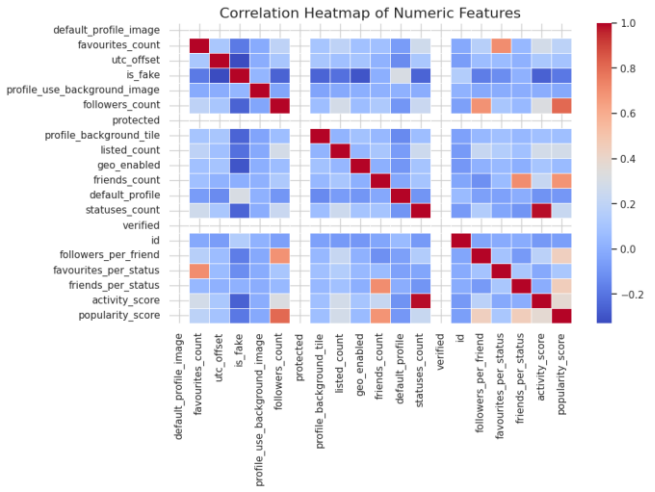



*Fig-4 Heatmap*

A correlation heatmap is a useful way to look at the interrelations of the numeric features in the dataset. The visualization with mostly light blue cells indicates that most features have weak correlations and hence the user behavior attributes provide unique and non-redundant information. However, the engineered features followers_per_friend, friends_per_status, activity_score, and popularity_score appear to be positively correlated among themselves to the extent of moderately strong, which means they are most likely capturing the same underlying characteristics with respect to user activity and engagement. The target variable is_fake is positively correlated to a significant extent with the metrics low follower count, high friends count, and activity-based ratios, thereby supporting the contention that fake accounts display and can be indirectly detected through their distinct and deep behavioral signatures. In sum, the heatmap not only confirms but also showcases the diversity of user behavior across different social platforms and the importance of the engineered features.

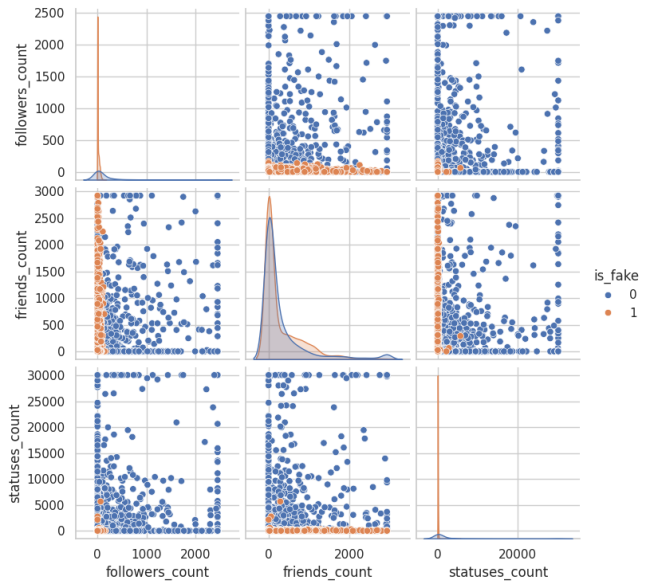



*Fig-5 Pairplot*

The pair plot showcases the correlation of key activity metrics which are the number of followers, the number of friends, and the number of statuses—for both real and fake users. A very noticeable pattern is seen: the fake accounts (marked with orange dots) are found at a very low level of engagement and activity as they are close to the bottom end of all three metrics. Conversely, the real users (marked with blue dots) show a much larger spread with greater variability in terms of followers, friends, and statuses. This difference implies that the accounts that are fake are usually devoid of natural growth and diversity in behavior. The diagonal density plots also corroborate this observation; they indicate that the maximums for the fraudulent accounts are nearly non-existent, whereas the authentic ones exhibit more extensive and organic distributions. In conclusion, this visual representation gives powerful evidence for the presence of a considerable difference between the genuine and the counterfeit profiles and, as a result, the usage of these characteristics as strong predictors in the detection model has been validated.

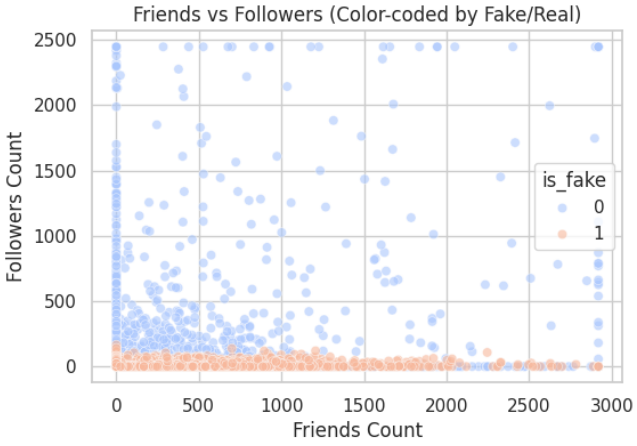



*Fig-6 Scatterplot*

The scatter plot showing the relationship between friends count and followers count gets a very clear line between authentic and inauthentic accounts. The inauthentic accounts (represented by orange points) are always located at the lower part of the plot, indicating their very low follower counts even though they often have high friends counts. This situation is typical for fake accounts, where the bots usually follow a lot of users without getting any engagement from them. On the other hand, the honest users (marked by blue points) occupy a much wider and more realistic area with various numbers of followers and friends that are horizontally distributed over the whole range. This stark contrast confirms the effectiveness of the ratios of followers to friends as a distinguishing factor in the identification of suspicious accounts.

These visual indicators were very much in tune with the model's learning patterns and made the difference between the classes quite obvious.

## V. DISCUSSION :

The results of this study highlight the effectiveness of classical machine learning techniques when combined with thoughtful preprocessing and behavior-based feature engineering for detecting fake social media accounts. The strong performance of the Random Forest and SVM models demonstrates that even without deep learning, robust patterns can be captured from user metadata and activity signals. The consistently high recall values across these models indicate their ability to correctly identify fake accounts, while high ROC–AUC scores confirm strong separability between real and fake users.

An important finding of the exploratory data analysis is that the behavior of fake accounts is inconsistent but recognizable. The use of visuals in the analysis indicates that fake profiles characteristically have very few followers, a lot of friends, and a very low posting frequency. Such characteristics are sometimes found in automatically created accounts or accounts with little effort put in. The heatmap showing correlation between the metrics confirms this behavior as it reveals that the metrics developed by the company such as followers_per_friend, activity_score, and popularity_score are significantly connected to the fake-account label. This strengthens the initial assumption that the engineered features of the close-up signal are very important in distinguishing the case from traditional raw features that are not able to capture the close-up signal at all.

The pair plots and scatter visuals offer deeper insight into the separation of the two classes. Fake accounts cluster tightly near the lower end of followers and statuses, while real accounts display broader, more natural distributions. This visual separability indicates that fake accounts are structurally different from genuine users in their activity patterns—a finding that supports the high performance of the classification models.

However, there are still some limitations that need to be addressed even though the results are good. The dataset has a natural imbalance with real accounts being more than fake ones. Even though the problem is considerably reduced by techniques such as stratified sampling and choosing evaluation metrics carefully, future research may try out advanced resampling techniques or anomaly-detection approaches to enhance the robustness of the model even more. Moreover, only numeric and behavioral metadata are available in the present dataset; the accuracy might be improved more by the integration of textual content, network-graph relationships, or temporal posting patterns, particularly for the advanced fake accounts that impersonate human behavior.

Overall, the findings emphasize that machine learning—supported by well-designed feature engineering and visual diagnostics—provides a scalable and reliable approach to identifying fake accounts. The results encourage the deployment of such models in real-world social platforms, where early detection can significantly reduce the spread of misinformation, spam, and malicious activity.

## VI. LIMITATION

Despite achieving strong performance, the proposed Fake Account Detection system has several limitations that must be acknowledged:

1. Dataset Imbalance:

The number of real accounts significantly outweighs the number of fake accounts, which may bias certain models despite stratification. This imbalance can reduce the system's sensitivity to rare but sophisticated fake profiles.

2. Limited Feature Scope:

The dataset primarily contains numeric and behavioral metadata. It does not include textual content (tweets, bios), interaction graphs, or temporal posting patterns, which could capture more complex fake-account behaviors that metadata alone cannot reveal.

3. Platform-Specific Behavior:

The features and patterns studied are tailored to the available dataset and may not generalize perfectly across different social media platforms with different user behavior norms or metadata structures.

4. Static Snapshot of Data:

The analysis is based on a frozen dataset and does not account for evolving behavioral tactics used by bot creators or fake account operators. Real-world fake accounts adapt quickly, potentially reducing long-term effectiveness.

5. Feature Engineering Dependency:

The system relies heavily on engineered features such as ratios and activity scores. If attackers deliberately mimic these patterns, the model may become less discriminative without additional deeper features.

6. Lack of Network-Graph Insights:

Fake accounts often interact in coordinated clusters. Since this system does not analyze social network graph structure (followers/following graph, retweet networks), it may miss detection of coordinated botnets.

7. Interpretability Challenges in Ensemble Models:

Despite Random Forest reaching the highest accuracy, its predictions are less understandable than those of simpler models, which may hinder real-world validation in high-stakes moderation environments with complicated scenario

## VII.CONCLUSION:

This study presents a machine learning–based framework for detecting fake social media accounts using behavioral metadata and engineered activity features. Through systematic preprocessing, feature engineering, visualization, and model evaluation, the system demonstrates strong capability in distinguishing real users from fake or automated profiles. The experimental results confirm that classical machine learning algorithms—particularly the Random Forest and SVM models—achieve high accuracy, recall, and ROC–AUC, proving their effectiveness for this task even without deep learning components.

The visual explorations also show that fake profiles have certain characteristics which can be easily spotted: they have few followers, they follow many people, and they do not post much. These features that have been expressed through measures and activity scores are helping a lot in increasing the discriminative power of the system. The correlation analysis affirms the significance of the engineered features and emphasizes their role in the performance of the model.

Despite promising results, the study recognizes several limitations, including dataset imbalance, the absence of textual or network-graph features, and the evolving nature of fake account behavior. Addressing these challenges opens opportunities for future work, such as integrating NLP-based content features, graph-based user interaction analysis, time-series posting patterns, or hybrid deep learning approaches.

Overall, this research demonstrates that an interpretable and computationally efficient machine learning pipeline can serve as a practical and scalable solution for detecting fake accounts on social media platforms. By strengthening content authenticity and reducing malicious activity, the proposed

system contributes to safer and more trustworthy online environments.

## VIII. REFERENCES


[1] C. M. Al-Qurishi, M. S. A. Khan, M. Gupta, M. Al-Rakhami, and A. Alamri, "A framework for identifying fake accounts on social networks," *IEEE Access*, vol. 7, pp. 65579–65589, 2019.

[2] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, no. 1, pp. 280–289, 2017.

[3] 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 *Proc. 26th Int. World Wide Web Conf.*, 2017, pp. 963–972.

[4] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 811–824, 2012.

[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Proc. CEAS Conf.*, 2010.

[6] M. Ahmed and A. Mahmood, "Anomaly detection in online social networks," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–18, 2014.

[7] A. K. Jain and B. B. Gupta, "Towards detection of malicious accounts on social networking websites," *Security and Communication Networks*, vol. 7, no. 2, pp. 526–540, 2014.

[8] J. E. Ho and S. P. Chew, "Fake account detection on social networks using machine learning techniques," in *Proc. IEEE ICCE*, 2020, pp. 457–462.

[9] A. Abokhodair, D. Yoo, and D. W. McDonald, Dissecting a social botnet: Growth, content, and influence in Twitter," in *Proc. 18th ACM CSCW Conf.*, 2015, pp. 839–851.

[10] T. Cresci, M. Petrocchi, and A. Spognardi, "Better safe than sorry: An adversarial approach to improve social bot detection," *Proc. WWW Companion*, pp. 137–138, 2018.

[11] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Proc. IEEE/ACM ASONAM*, pp. 551–554, 2018.

[12] H. M. J. Almuhammadi and M. Alsulami, "Detecting fake social media accounts using machine learning and behavioral analytics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 499–506, 2020.

[13] G. Wang, "Social media bot detection via behavioral feature engineering," *J. Inf. Secur. Appl.*, vol. 58, 2021.

[14] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Anomaly detection in online platforms using metadata and behavioral signatures," *IEEE BigData*, pp. 2273–2280, 2019.