Original Research Article

Detecting Fake Accounts on Instagram using Machine Learning

Nikita Efimov

The Woodlands High School, The Woodlands, TX 77381, United States

ABSTRACT

The existence of fake users on social media platforms like Instagram creates significant challenges for marketers and platform integrity. Fake users are usually used for engagement manipulation such as spamming and fake followings. This study investigates the problem of identifying fake users using machine learning techniques, leveraging rich metadata from a dataset with 65326 Instagram accounts. Using analysis of 18 metadata features, two models -Random Forest and XGBoost-were trained and evaluated using precision, recall and F1-score. XGBoost achieved the best performance, with F1-score of 0.91 for real users and 0.9 for fake users. Feature importance analysis using SHAP, underlined link availability, engagement rate (comments), and engagement rate (likes) as the most predictive features. This study underscores the potential of machine learning in combating fake user expansion and provides insights for improving model interpretability and efficiency. Future research could explore larger datasets, integrate more advanced techniques and incorporate additional metadata to further refine detection models.

Keywords: Fake users; Social media; Instagram; Machine learning; Metadata; Engagement manipulation; Spam detection; Fake followers

INTRODUCTION

Instagram is a popular social network with millions of active users. Often, marketing agencies use Instagram influencers with large numbers of followers to promote their products, but sometimes those followers are fake. Usually, fake users are considered to be bots; however, a report found that some real users sell their passwords, therefore expanding the definition of fake user. Fake followers can make up a significant number of user's followers. High follower counts can mislead companies to pay more for promotion of their product, despite fake users sometimes making up to 78% of followers (1).

Identifying fake users is a pressing problem on social media platforms. Many studies used Twitter and LinkedIn with limited metadata and supervised machine learning techniques to detect fake users from real users (2). But Instagram or Meta platforms contain richer metadata about the posts and user accounts. Very important aspects that were used to determine fakeness of users are similarities, such as names, post frequency and date of account creation. These aspects can be used to determine user status.

A study was conducted by Purba to identify fake or real users using Instagram data (3). Their best model achieved precision, recall and f1 scores of 90 percent. They showed that the descriptive statistics

Corresponding author: Nikita Efimov, E-mail: nikita.s.efimov@gmail.com.

Accepted October 6, 2025
https://doi.org/10.70251/HYJR2348.35710716

helped to differentiate between fake and real users (**Table 1**). Similar research was done by Ümmü Tunç *et al*. (4). However, in their study they used only 2799 accounts which creates under representational bias. The objectives of this study are (1) to develop and evaluate machine learning model which identifies if an Instagram user is fake or real with higher accuracy and (2) understand and quantify which factors or metadata contribute to model predictions.

METHODS AND MATERIALS

Dataset

Purba *et al*.'s (2019) dataset about fake and authentic Instagram was used. The data was collected from September 1st, 2019, to September 20th, 2019[3]. The target column was a class which had two outputs, such as real and fake. There are 32460 real users and 32866 fake users. There are 18 columns which contain criteria by which authenticity was judged such as number of posts, number of followers, number of people whose user is following, biography length (Table 1).

Exploratory Data Analysis

In this section findings from exploratory data analysis are presented. The goal is to understand the patterns and the correlation between the features, obtain valuable insights and trends in the dataset. To better understand inter-feature relationships, a Spearman correlation heatmap was generated to visualize how strongly the features were associated with one another. As shown in Figure 1, the dataset does not contain any features with correlations above 0.65, indicating minimal collinearity and confirming its suitability for machine-learning model training.

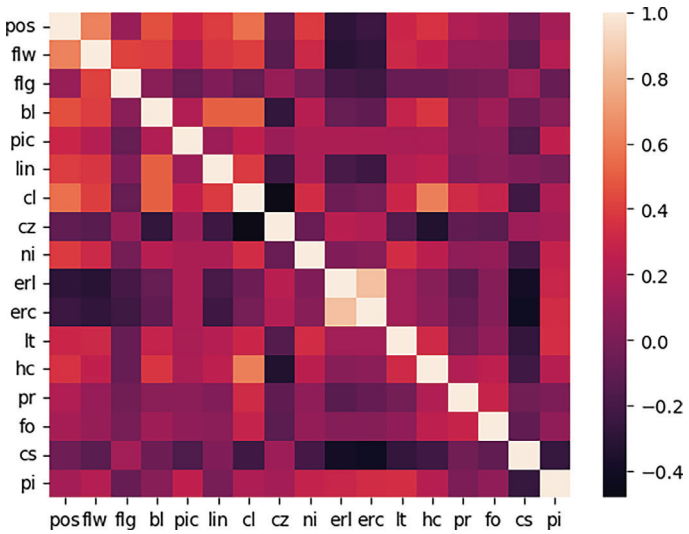

Figure 1. Correlation matrix illustrates spearman correlation between the features. This heat map clearly shows that this dataset does not contain any features with correlation above 0.65. This criterion is crucial for model development because strong correlation between features is the sign of collinearity.

Table 1. Dataset features used for classification, including user activity, engagement, and profile metadata

Feature	Description
pos	Num posts| Number of posts that user posted
flw	Num following| Number of followings that user has
flr	Num flr | Number of followers that user has
bl	biography length| Length (in characters) of user's biography
pic	Picture availability | 0 if user doesn't have it 1 if does
lin	Link availability |1 if user has link available 0 if not
cl	Average caption length| average number of characters of caption in user's media
cz	caption zero | (0 .0 to 1.0) percent of captions with almost 0 length
ni	Non image percentage| percentage of non-image media
erl	engagement rate (like)| Number of likes divided by number of followers
erc	engagement rate (comments)| Number of comments divided by number of followers
lt	location tag percentage| percentage of posts with location
hc	hashtag count| Average number of hashtags in user's posts
pr	promotional keywords| Average number of promotional words
fo	Followers' keywords|average use of hunter words, hashtags etc
cs	Cosine similarity|Average cosine similarity between posts that user has
pi	post interval| Average post interval (in hours)

To further explore content-related behavior, the distribution of near-empty captions ("caption zero") between real and fake users was analyzed. Figure 2 illustrates that fake accounts tend to post more frequently with near-empty captions, whereas genuine users usually include descriptive or meaningful text. This pattern suggests that caption completeness can serve as a strong authenticity indicator.

As shown in Figure 3, fake users tend to follow a much larger number of accounts (median ≈ 2,000), while genuine users follow significantly fewer (median ≈ 500). This supports the hypothesis that artificially boosted engagement patterns characterize fake profiles.

The follower-count distribution was also examined to complement these findings. As shown in Figure 4, fake accounts typically possess fewer followers (median ≈ 200) compared to real users (median ≈ 400), reinforcing the notion that follower–following imbalance is a distinguishing feature of inauthentic behavior.

The role of external links in user profiles was evaluated next. Figure 5 demonstrates that real users more frequently include active links in their bios,

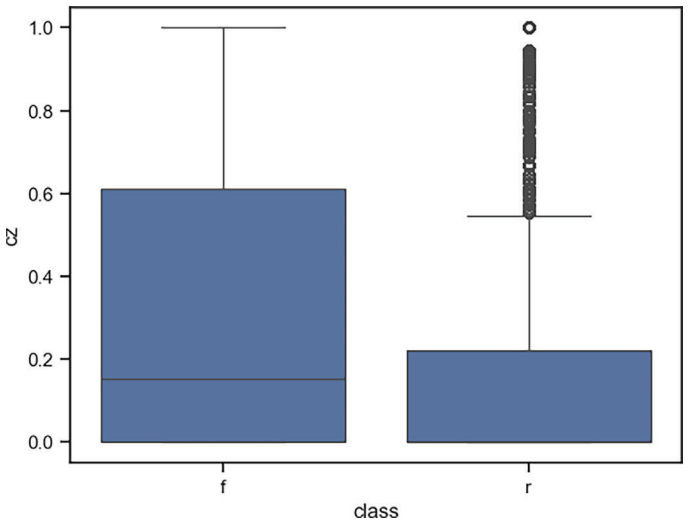

Figure 2. Comparison of caption zero of fake users to real users.

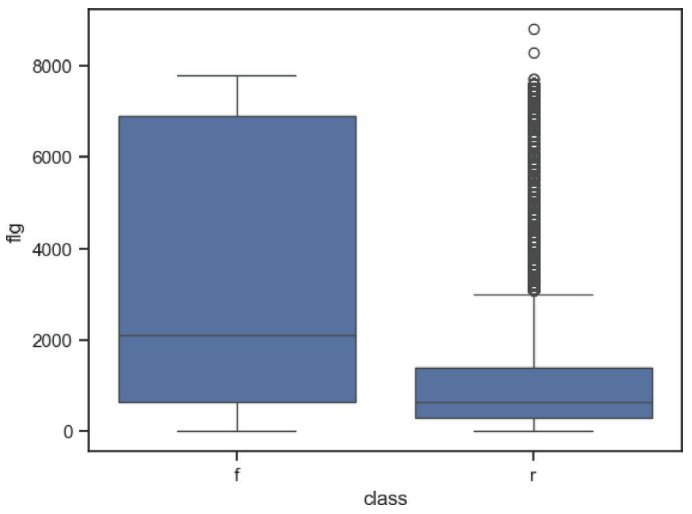

Figure 3. Number of followers that user has for real and fake accounts boxplot.

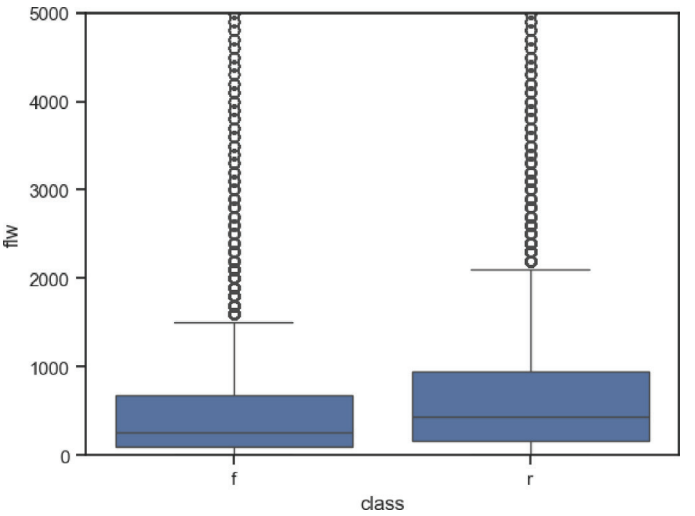

Figure 4. The figure shows the boxplot of number of followers that user has. As shown on the boxplot, on average, fake users tend to have less followers (median=200) than the real ones (median=400).

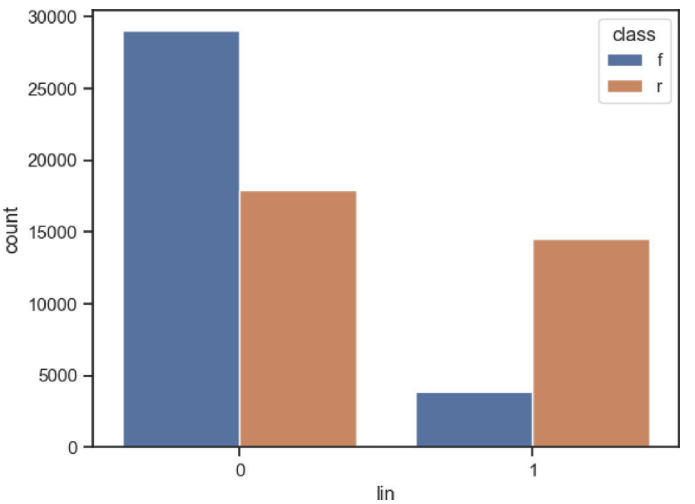

Figure 5. The figure shows the count distribution of link availability for real and fake users.

whereas fake users often omit them. This distinction later emerged in SHAP analysis as one of the most influential predictors for authenticity classification.

Engagement dynamics were also analyzed to determine behavioral differences between real and fake users. Figure 6 presents boxplots of engagement rates based on likes, showing that fake accounts generally receive lower engagement per follower and exhibit higher variability, possibly due to automated interactions.

Similarly, Figure 7 illustrates engagement rate distributions measured by comments, where genuine users exhibit higher and more consistent comment activity. Together, these findings highlight engagement metrics as strong discriminative features for model training.

Models and Evaluation Metrics

In this study, the random forest and XGboost models are implemented to identify fake users from real users (Table 2). The random forest model is a machine learning technique that combines multiple decision trees to improve accuracy of predictions. It works by creating a "forest" of decision trees, each trained on its own part of data, and makes predictions based on the average prediction (5).

XGBoost is a machine learning algorithm that is based on gradient boosting, which builds a series of trees, where each of them is trying to correct errors of previous one. It is also popular for its ability to handle big datasets, and it is highly customizable which allows users to tune parameters (6).

To evaluate the best model precision, recall, F1-score are used (Table 2, Table 3). Precision measures the accuracy of positive predictions. It is defined as the ratio of predicted positives divided into the number of total positives. Recall measures how good the model identifies all relative instances. It can be defined as the number of true positives divided by the number of false negatives and true positives. The F1-score basically means precision and recall. Those measures are used to determine the effectiveness of the model. When it comes to the evaluation of comparison of training and testing results it is important to check if the model performs similarly in both sets. Significant differences between training and test scores can indicate that the model is either underfitting or overfitting, suggesting that the model is too specialized for training data.

Hyperparameter Tuning

Hyperparameter tuning is one of the most important steps in model development. Hyperparameters (HP)

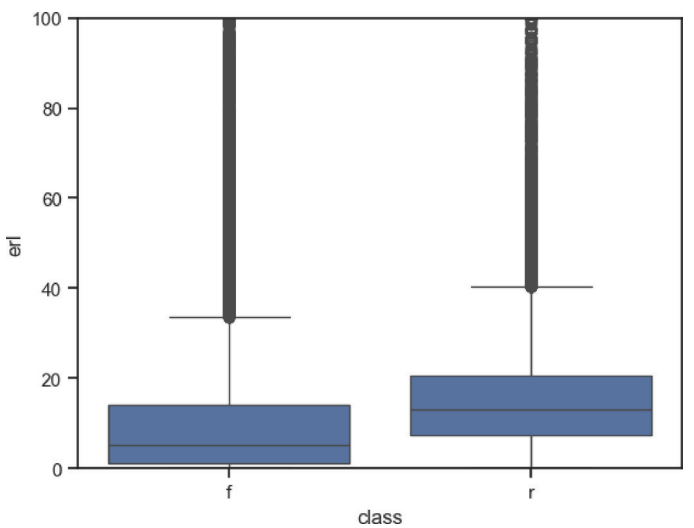

Figure 6. Engagement rate (like) for real and fake users' boxplot. It illustrates that the majority of fake users have an engagement rate between 0 and 15 with a noticeable number of outliers, while real users have an engagement rate ranging from 10 to 20 and a large number of outliers.

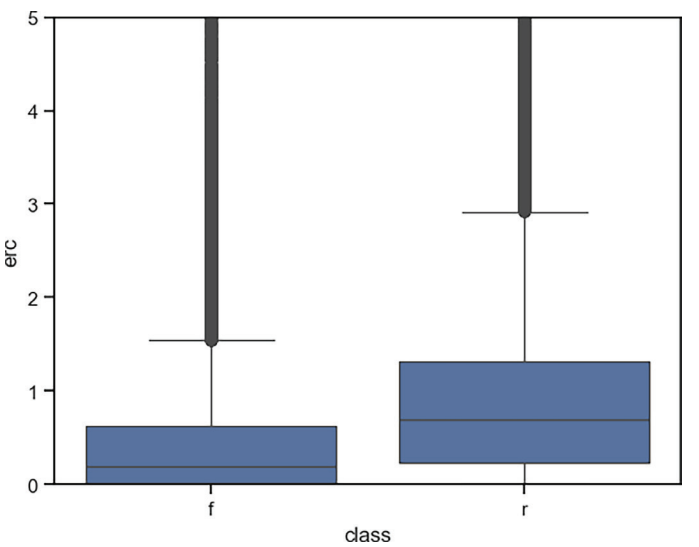

Figure 7. Engagement rate (comment) for real and fake users. The engagement rate, measured by comments, most of the fake users have an engagement rate close to 0.2, while real users have an engagement rate of 0.5. Both fake and real users have a large number of outliers, suggesting that some users have different patterns of behavior.

Table 2. Performance comparison of Random Forest and XGBoost models, showing precision, recall, and F1 scores for fake and real users across training and test sets

Model	Train/Test	Label	Precision	Recall	F1-score
random forest (max_features=9, min_samples_ spilt=40, n_estimators 500)	Train	real	0.89	0.98	0.94
		fake	0.98	0.88	0.93
	Test	real	0.86	0.96	0.91
		fake	0.96	0.84	0.90
XGBoost ((colsample_bytree 0.8, eta=0.05, n_estimators 500, subsample=0.8)	Train	real	0.91	0.98	0.94
		fake	0.98	0.90	0.94
	Test	real	0.87	0.95	0.91
		fake	0.95	0.86	0.90

Table 3. Performance of XGBoost using only the top 5 SHAP-identified features, compared to the full feature set

Model	Train/Test	Label	Precision	Recall	F1-score
XGBoost ((colsamble_bytr ee 0.8, eta=0.05, n_estimators 500, subsample=0.8)	Train	fake	0.96	0.82	0.88
		real	0.84	0.97	0.90
	Test	fake	0.96	0.82	0.88
		real	0.84	0.96	0.9

This highlights the trade-off between interpretability and accuracy.

are the external configurations set before the learning process starts. They significantly affect the model's ability to make accurate predictions. For tree-based models such as Random Forest or XGBoost, several hyperparameters can be adjusted to optimize performance.

For the random forest model following parameters were used with following ranges. n_estimators (number of trees) had range from 500 to 1000; min_sample_split - minimum number of samples needed to create split had range from 40 to 70 with increment of 20. max_ features maximum number of features used had range from 1 to 17 with increment of 2.

For the XGBoost model hyperparameters with following ranges were used: n_estimators is the number of boosting rounds that range from 100 to 500. learning_ rate(eta) is important for controlling the contribution of each tree ranging from 0.01 to 0.3. The subsample determines the fraction of features used to train each tree ranging from 0.5 to 0.9. Colsample_bytree, a parameter which controls the fraction of features used to create a tree, had a range from 0.5 to 0.9.

Feature importance analysis

In this study SHAP analysis implemented to understand and quantify marginal contribution of individual features on model predictions (Figure 8, Figure 9, Table 1). SHAP is an important machine learning explainability technique based on cooperative game theory, which works by calculating each player's contribution to the final outcome (7). Similar to cooperative games, SHAP explains how much each feature contributed to the final prediction. SHAP values are model agnostic, meaning that they can be used to explain machine learning models such as linear regression, decision trees, random forest and XGBoost.

RESULTS AND DISCUSSION

The results for the Random Forest model indicate strong performance (Table 2). In the training set, it achieved a precision of 0.89, recall of 0.98, and F1 score of 0.94 for real users, while for fake users, the precision was 0.98, recall 0.88, and F1 score 0.93. On the test set, the precision, recall, and F1 scores for real

users were 0.86, 0.96, and 0.91, respectively, and for fake users, they were 0.96, 0.84, and 0.90. Similarly, the XGBoost model showed a good performance (Table 2). For the training set, it achieved a precision of 0.91, recall of 0.98, and F1 score of 0.94 for real users, and for fake users, it achieved a precision of 0.98, recall of 0.90, and F1 score of 0.94. In the test set, the precision, recall, and F1 scores for real users were 0.87, 0.95, and 0.91, respectively, and for fake users, they were 0.95, 0.86, and 0.90. Overall, both models showed strong performance, with XGBoost slightly overperforming Random Forest. Because XGBoost over-performed Random Forest it was decided to choose XGBoost to identify parameters needed that contribute to model predictions.

To interpret the XGBoost model's predictions, SHAP analysis was applied to quantify each feature's marginal contribution. The SHAP summary plot displays the overall influence of each feature across all instances, where features such as link availability and engagement rate (comments) show large positive SHAP values for authentic users (Figure 8).

Figure 9 summarizes the average SHAP impact values for all features, ranking them by importance. The top predictors—link availability, engagement rate (comments), engagement rate (likes), number of followings, and number of followers—demonstrate the model's ability to prioritize social-behavioral indicators over superficial profile attributes.

After SHAP analysis, XGBoost model was hyperparameter tuned with only top 5 contributing features, link availability, engagement rate (comment), engagement rate (like), number of followings and

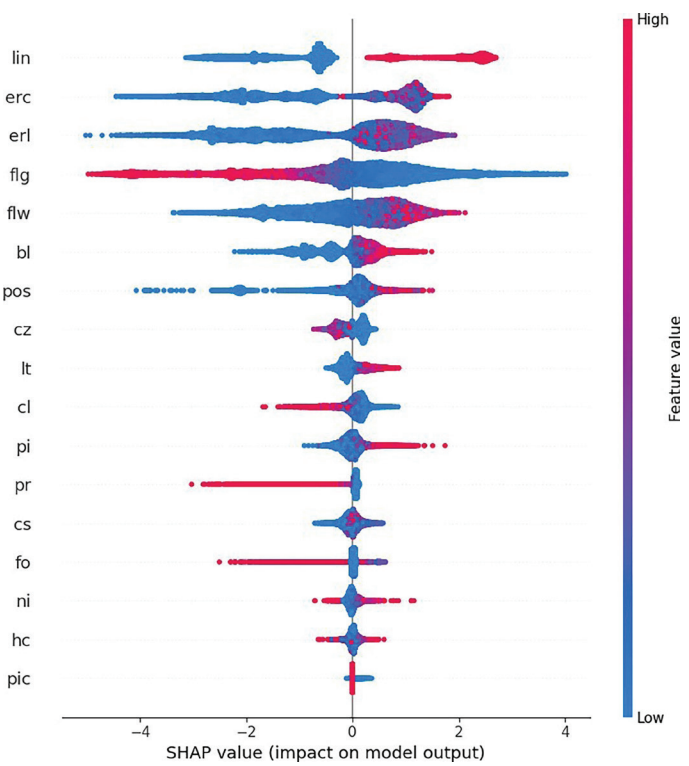

Figure 8. SHAP summary plot, which shows marginal contribution of each individual feature on model predictions. Each dot represents a single observation in the dataset, with its horizontal position indicating SHAP value. Features are shown in descending order of significance from top to bottom. The color of each dot represents the feature value, with red representing high values and blue representing low values.

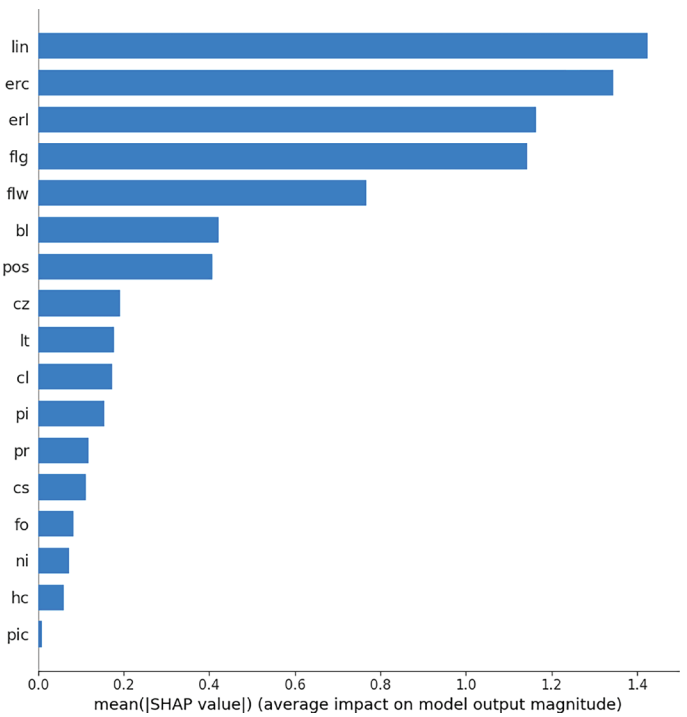

Figure 9. A bar chart summarizing the average impact of features on model predictions based on the mean absolute SHAP values. The x-axis represents the average SHAP value, while features are arranged in descending order of importance from top to bottom. Key features such as link availability, engagement rate (comments), and engagement rate (likes) show the highest average influence, highlighting their critical role in the model's decision-making process. In contrast, features like hashtag count and picture availability have minimal contributions, indicating their relatively lower relevance to the predictions.

number of followers. However, it was found that for fake users, precision decreased from 0.98 to 0.96 and the F1-score decreased from 0.94 to 0.88. For real users, F1-score decreased from 0.94 to 0.90, recall decreased from 0.98 to 0.97 (Table 3).

For the test data, model with evaluated features had slightly better precision for fake users (0.96 and 0.95 respectively) but recall slightly decreased from 0.86 to 0.82 and F1 score decreased from 0.90 to 0.91. For real user tests both models had the same recall of 0.95 and F1 score decreased from 0.91 to 0.9 (Table 3).

Overall, the model with top contributing features demonstrated a marginally worse performance than model without evaluated features.

CONCLUSIONS

The outcome of this research provides a foundation for future research in detecting fake users on Instagram. It demonstrates the effectiveness of machine learning models in dealing with this challenge and emphasizes the importance of using metadata to improve prediction accuracy. Future studies could focus on expanding the dataset, integrating more advanced deep learning techniques such as Attention- Based LSTM, Bidirectional GRU, Bidirectional LSTM, LSTM with Flatten Layer, and Bidirectional LSTM, which were used in Diptam Mukhopadhyay's research, and exploring additional metadata to further increase model capabilities (11). The findings of this study have practical implications for social media platforms and marketers in identifying and reducing the influence of fake accounts, ensuring more authentic interactions and engagements. The dataset analysis helped evaluate features important to determine if the user is fake or real. Using SHAP analysis, it was found that features that contribute most to the model predictions are link availability, engagement rate (comment), engagement rate (like), number of followings and number of followers.

ACKNOWLEDGEMENTS

The author expresses his greatest appreciation to Abdulla Kerimov for his patience, guidance and knowledge shared with him.

CONFLICT OF INTERESTS

The author declares no conflicts of interest related to this work.

REFERENCES

1. Neff J. *Study of influencer spenders finds big names, lots of fake followers*. AdAge. 2018, April 23. https://adage.com/article/digital/study-influencer-spenders-finds-big-names-fake- followers/313223/ (accessed on 2025-07-21)
2. Ramalingam D & Chinnaiah V. Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering,* 2017; 65*:* 165–177. https://doi.org/10.1016/j.compeleceng.2017.08.033
3. Purba KR, Asirvatham D & Murugesan RK. Identifying fake users on Instagram using metadata and machine learning algorithms. *Computers & Electrical Engineering.* 2019; 76: 91–102. https://doi.org/10.1016/j.compeleceng.2019.03.017
4. Tunç Ü, Atalar E, Gargı MS & Aydın ZE. Classification of fake, bot, and real accounts on Instagram using machine learning. *Politeknik Dergisi,* 2024; 27 (2): 479–488. https://dergipark.org.tr/en/download/article-file/2509052
5. Baladram S. *Random forest | Towards Data Science*. Medium. 2024, November 30. https://medium.com/@sbaladram/random-forest-tutorial (accessed on 2025-06-15)
6. Calderon J, Jr. *What is XGBoost?* Medium. 2023, October 7. https://medium.com/@xthemadgenius/what-is-xgboost-cf6650345edd (accessed on 2025-08- 01)
7. Savcı U. *Complete exploratory data analysis using Python*. Medium. 2022, April 3. https://medium.com/@ugursavci/complete-exploratory-data-analysis-using-python- 9f685d67d1e4 (accessed on 2025-07-11)
8. Gurajala S, White JS, Hudson B, Voter BR & Matthews JN. Profile characteristics of fake Twitter accounts. *Big Data & Society.* 2016; 3 (2): 1–12. https://doi.org/10.1177/2053951716674236
9. Sharma A, Agarwal R, Singh M & Pant A. Fake profile detection on social networking websites: A comprehensive review. *IEEE Access.* 2020; 8: 212865–212889. https://doi.org/10.1109/ACCESS.2020.3040195
10. Gupta P, Ranganath S & Suma V. Worth its weight in likes: Towards detecting fake likes on Instagram. *ResearchGate. 2018.* https://www.researchgate.net/publication/325211150_Worth_its_Weight_in_Likes_Towards_Detecting_Fake_Likes_on_Instagram (accessed on 2025-07-05)
11. Mukhopadhyay D, Chowdhury A, Sarkar S, Goenka P, *et al*. Enhancing fake account detection on digital platforms using deep learning models. In *Proceedings of the 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT).* 2025;pp.153–158.IEEE.https://doi.org/10.1109/CSNT64827.2025.10968093