[[1]](#footnote-1)

DAT 554 Final Project:

Fake Instagram account binary classification using machine learning

Carissa Hicks  
*University of Illinois at Springfield, Data Analytics*

*Abstract*—Fake social media accounts and social media engagement can be incredibly dangerous. Illegitimate social media accounts can be used for a variety of purposes including increasing popularity of a main account to join an affiliate program, emotional scams, or phishing. The purpose of this study is to explore a dataset consisting of 994 real accounts and 200 fake accounts for features relating to publicly available information. Machine learning algorithms such as Naïve Bayes, Neural Networks, and Logistic Regression will be applied to build a model to detect these accounts.

***Keywords - machine learning, Instagram, social media, fake accounts, binary classification***

# INTRODUCTION

S

ocial media is a very important part of many people’s daily lives. It is where business market their products, influencers interact with their fans, and where family and friends share and like posts amongst each other. Some of the most notable social media application include Facebook, Twitter, Instagram, and YouTube [10]. However, as the importance of keeping up with relevant people, businesses, and happenings increases, so do methods of taking advantage of these systems.

Instagram is a free photo and video sharing app where a user can follow people, gain followers, share, comment and like posts. The popularity of an Instagram account can be measured by account engagement. Engagement includes total followers, total posts, total comments on each post, and total likes on each post. [11] Fake accounts to increase these metrics can be created using automated generators. These accounts can be programmed to like posts, follow accounts, and even comment and share posts automatically to increase the account’s engagement. The popularity and even reliability of an account depends on these metrics. A popular account is more likely to be trusted. These accounts can engage in Sybil attacks (operating many fake identities simultaneously), scams, or sending malicious links and spam to other users. The detection of these kinds of accounts is vital to avoid these events.

It is important to note that there is a distinction between the usage of creating fake accounts for identity theft, and accounts used solely for increasing metrics and spam. There would be more effort going into an account that wants to impersonate an individual such as a celebrity or target’s family member. These accounts would include a profile picture and style their username to match the impersonated individual’s real username as closely as possible. However, these accounts can still be detected by their follower count and number of posts.

Instagram has already introduced measures to combat against bots and scammers. “We’ve built machine learning tools to help identify accounts that use these services and remove the inauthentic activity.” [19] Although, services that offer Instagram automation are publicly available and easy to use. Colleen Christison, a writer at Hootsuite, demonstrated various apps available that use Instagram automation [8]. In recent years, several research articles have been released introducing fake account detection. Akyon et al [2] provided the dataset that was utilized in this analysis. The authors make a distinction between fake and automated accounts and were able to classify them with 86% accuracy using a neural network model.

Sheikhi et al [21] introduced a machine learning model based on features such as username length, profile picture, the number of followers vs following, if it is a public or private account, post count, and others. They found that most accounts that do not have a profile picture and/or are following more people than they have followers, are usually fake accounts. They were able to obtain a 98% classification accuracy using bagged decision trees.

This paper will answer the following questions: What are the characteristics of a fake account? Can fake accounts be detected with publicly available data? Do these features apply to other social media platforms?

# Dataset

The dataset was obtained from GitHub uploaded by Faith Cagatay Akyon. The author of the dataset along with Esat Kalgaoglu also provided their own analysis of the data. This data was hand labeled by the authors using publicly available information. It is important to note that none of the data is related to the contents of each profile. Private accounts do not show the personal information of the profile such as their posts or comments. The goal of this analysis is to analyze data that anyone can obtain.

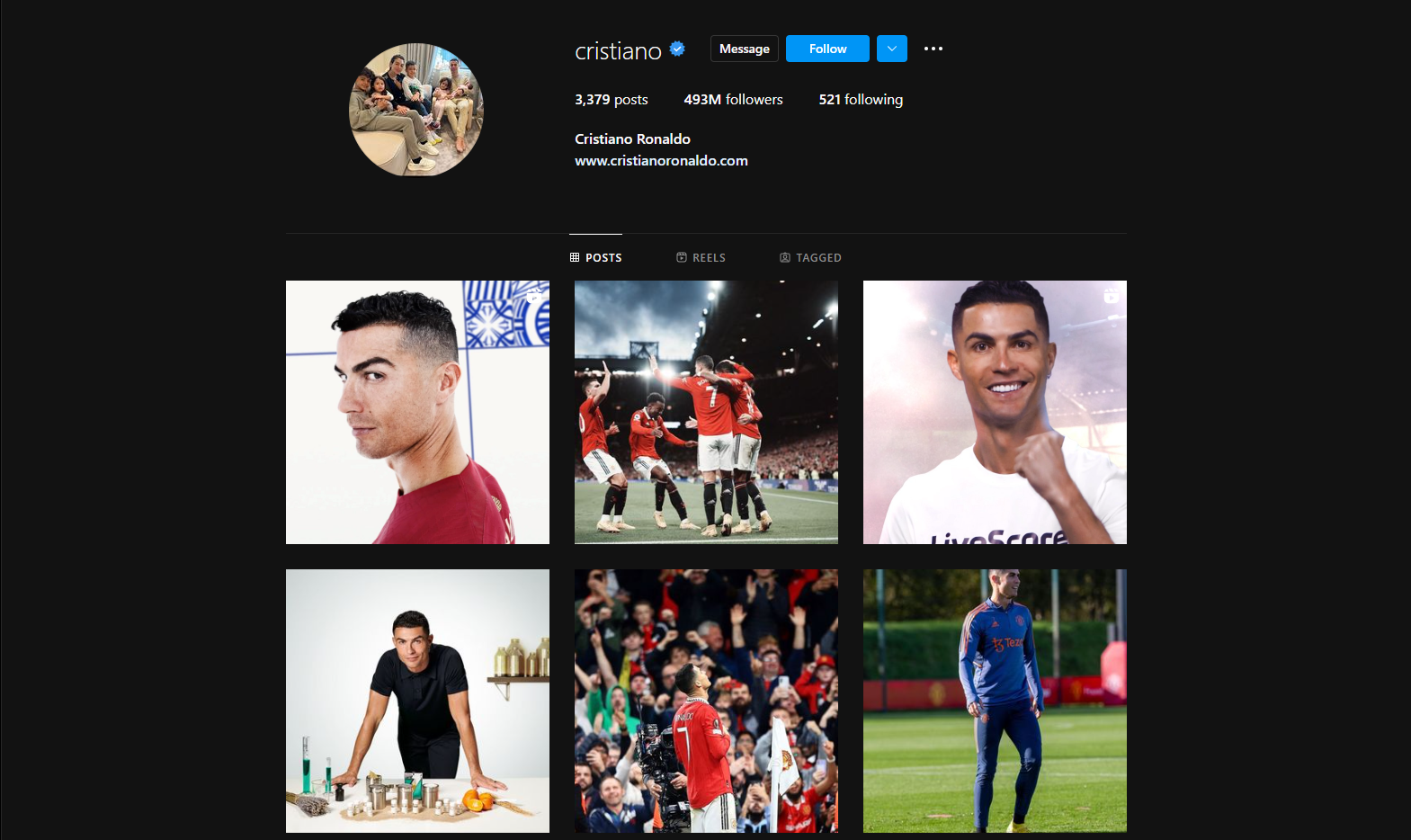


Figure - Example of a public Instagram account

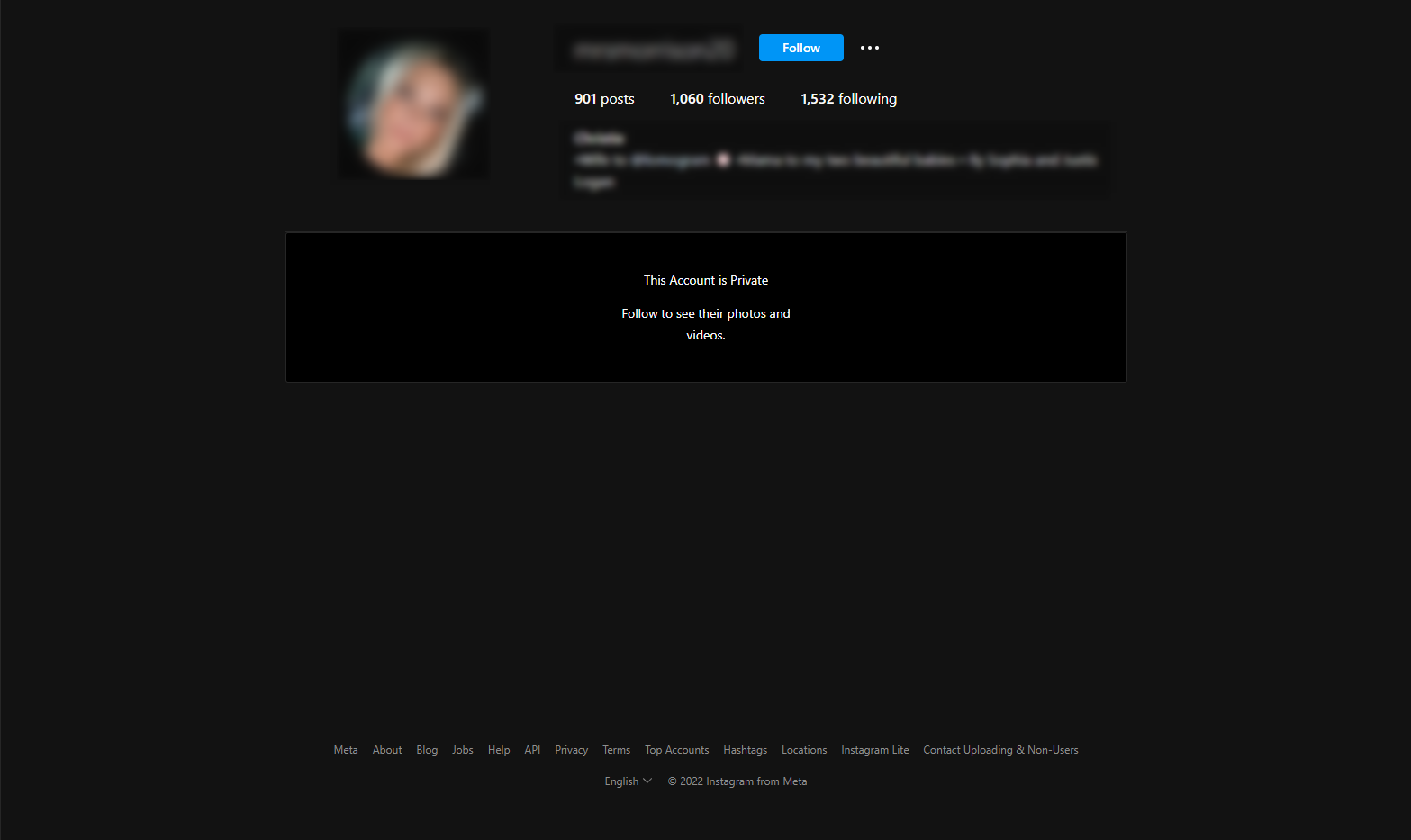


Figure - Example of a private Instagram account

The original package contained csv files of fake accounts and real accounts separated. The fake account file contained 200 observations of 9 variables while the real account file contained 994 observations of 9 variables. The 9 variables are as follows:

* isFake: is the account fake or real (binary)
* userBiographyLength: length of characters in the biography
* userFollowerCount: number of users the account follows
* userFollowingCount: number of users the account is following
* userHasProfilPic: does the account have a profile picture (binary)
* userIsPrivate: is the account private (binary)
* userMediaCount: number of media the account has uploaded
* usernameDigitCount: number of digits in the username
* usernameLength: length of the username

During preprocessing of the dataset, the two separate files were combined into one dataset of 1194 observations. The data did not contain any missing variables or null observations. For initial exploration of the data, isFake, userHasProfilPic, and iserIsPrivate were converted to factors since they are binary categorical variables. Then, real/fake was renamed as factor levels to or yes/no respectively.

Chart, bar chart

Description automatically generated

Figure - Data distribution of private vs public accounts

Chart, bar chart

Description automatically generated

Figure - Data distribution of "profile picture" feature

Chart, box and whisker chart

Description automatically generated

Figure - Data distribution of biography lengths

Chart

Description automatically generated with medium confidence

Figure - Data distribution of follower counts

Chart, box and whisker chart

Description automatically generated

Figure - Data distribution of following counts

A picture containing chart

Description automatically generated

Figure - Data distribution of media counts

Chart

Description automatically generated

Figure - Data distribution of username digit counts

Chart, box and whisker chart

Description automatically generated

Figure - Distribution of username lengths

Furthermore, an additional helpful feature was added that was derived from userFollowerCount and userFollowingCount called followRatio. Instagram does not allow users to follow more than 7,500 accounts [13]. Fake accounts will near this number to increase their chances of getting a “follow-back” which will increase their stats [11] .

**There is an issue when the userFollowerCount or userFollowingCount is zero. If we divide by zero it results in an undefined value. To remedy this issue, all instances in which the counts equal zero will be changed to 1 for the followRatio calculation to take place. A higher ratio indicates that the account has more followers than they are following. The distribution of our follow ratio data indicated that fake accounts have a much lower follow ratio than real accounts as shown in the below figure with the mean indicated with the dotted black line.**

Chart, histogram

Description automatically generated

Figure – Follow ratios

One issue with the data is that there are mostly real accounts versus fake accounts which may throw off our prediction training. The majority class in this case is real accounts and the minority class are fake accounts. About **83.25%** of the data consists of real accounts whereas **16.75%** of the data consists of fake accounts. This analysis is interested in predicting the minority class – fake accounts.

Chart, bar chart

Description automatically generated

Figure - Bar chart demonstrating the class imbalance

There are several methods to dealing with imbalanced data such as over-sampling the minority class, under-sampling the majority class, and generating synthetic data. Due to the reason that we have a relatively low amount of data in general (under 2000 observations) I will choose to generate synthetic data using the ADASYN. [1] ADASYN stands for Adaptive Synthetic algorithm which creates synthetic instances of the minority class to achieve a better classifier performance. ADASYN will be applied after splitting the data into train/validation and test sets and only applied on the training set.

Further data exploration was conducted comparing the relationships between the target variable “isFake” with the other variables in the dataset. Categorical variables were explored with chi-squared test and mosaic plot. Using a confidence level of 95%, an alpha of 5% was used for the duration of this analysis. The two categorical variables in the data are userHasProfilPic and userIsPrivate. Both resulted in p-values lower than our alpha and are therefore, relevant to our analysis. The rest of the features were numerical and were analyzed using t-tests. The only variable with a p-value higher than alpha was usernameLength. Referencing back to Fig. 10 it is shown that there is almost no distinguishable difference between the boxplots of real and fake account username lengths. As a result, usernameLength will be removed from our data before building the machine learning models.

# Solution

First, the rows of the data frame should be shuffled randomly before splitting the data into test, train, and validation sets because the separate real and fake datasets were combined and may show a data separation bias. The data will be split into 80% training, 10% testing, and 10% validation sets. Models used include logistic regression, simple support vector machine, random forest, neural network, and naïve bayes. Then, ADASYN, feature scaling using min-max normalization, and cross validation was applied on the training data for relevant models. Min-max normalization was used instead of R’s default normalization method which is z-score normalization.

Min-Max Normalization

All models chosen are typically used for classification problems. Our baseline model, logistic regression was used to contextualize the results of the other trained models. The following is a list of the models used:

1. Logistic regression (baseline)
2. LR with adaptive synthetic minority oversampling technique on the fake accounts
3. LR, ADASYN, and cross validation
4. LR, ADASYN, CV, and min-max normalization
5. Simple support vector machine with normalization
6. Simple support vector machine without normalization
7. Random forest with ADASYN
8. Naïve Bayes with ADASYN and normalization
9. Naïve Bayes, ADASYN, and without normalization
10. Neural Network
    1. 32 units in the input layer (relu activation)
    2. 32 units in the hidden layer (relu activation)
    3. 1 unit in the output layer (sigmoid activation)
    4. Adam optimization
    5. Batch size = 32
    6. Epochs = 500

Of all these models, the logistic regression model with ADASYN, CV, and normalization performed the worse. This could be because of the feature scaling technique that was used. Referencing back to the boxplots created in figures 3-10 it is shown that the data contains a high number of outliers. This could be due to celebrity accounts that may have thousands of followers but zero following. These types of accounts are not typical of a normal everyday user’s Instagram profile. Although the min-max normalization algorithm performs better than z-score normalization overall it does have the downfall of not handling outliers very well [12]. If this project were to be redone, it would be interesting to see if z-score normalization produced better results.

# Evaluation Metrics

A picture containing calendar

Description automatically generatedEach model was tested on the test data and evaluated using accuracy, area under the curve, precision, recall, and F1. Since our data is imbalanced, more importance was placed on the F1 score because F1 score combines precision and recall into a single metric by using their harmonic mean.

Figure - Results chart

However, the logistic regression model with ADASYN scored the highest amongst all other models in every metric. The next best model would be the random forest model, then Naïve Bayes without normalization. It is interesting to note that the Neural Network model produced a recall of 1 indicating that there were no false negatives. However, it also has an AUC of 0.5 meaning that it is making random guesses for predictions and has no class separation capacity. This could indicate that the model needs more training. A future project may want to gather much more data to train the neural network to be a relevant model for predictions.

# Conclusion

This paper discussed the detection of fake accounts used on Instagram and the dangers they pose to social media users. Several machine learning algorithms, along with min-max feature scaling, and ADASYN algorithm was used to predict these fake accounts given publicly available features found on a user’s profile. As a result, logistic regression with adaptive synthetic minority oversampling produced the best F1-score for detecting fake accounts at about 97%. In future work, more data could be collected to increase the complexity of the models. The small amount of data available in the dataset obtained lackluster for training. In addition, with the inclusion of more data, outliers could be erased to give better representation to the general public. Also, because of these outliers, min-max normalization could be swapped out with z-score normalization which has been seen to work better with outliers.

Acknowledgment

I would like to express my acknowledgement to the professors at the University of Illinois at Springfield for their guidance during this project. Specifically professor Liang Kong who led the capstone course for this project, and professor Elham Buxton who I asked for support.

References

[1] *ADASYN — Version 0.9.1*. (n.d.). Imbalanced Learn. <https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html>

[2] Akyon, F. C., & Esat Kalfaoglu, M. (2019). InstaFake Dataset [Dataset; Python]. In *GitHub*. <https://github.com/fcakyon/instafake-dataset>

[3] Akyon, F. C., & Esat Kalfaoglu, M. (2019). Instagram Fake and Automated Account Detection. *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. <https://doi.org/10.1109/asyu48272.2019.8946437>

[4] Bhargava, K. (2021, December 24). *Instagram Fake Account Detection using Machine Learning*. Medium. <https://bhargavakapil24.medium.com/instagram-fake-account-detection-using-machine-learning-fd070f58b8f>

[5] Brownlee, J. (2020, January 14). *A gentle introduction to imbalanced classification*. Machine Learning Mastery. <https://machinelearningmastery.com/what-is-imbalanced-classification/>

[6] Chang, W. (2022, November 29). *R Graphics Cookbook, 2nd edition*. <https://r-graphics.org/>

[7] Chawla, N. V. (2002, June 1). *SMOTE: Synthetic Minority Over-sampling Technique | Journal of Artificial Intelligence Research*. <https://www.jair.org/index.php/jair/article/view/10302>

[8] Christison, C. (2022, September 16). *I Tried Instagram Automation (So You Don’t Have To): An Experiment*. Social Media Marketing & Management Dashboard. <https://blog.hootsuite.com/i-tried-instagram-automation-so-you-dont-have-to/>

[9] Ersahin, B., Aktas, O., Kilinc, D., & Akyol, C. (2017). Twitter fake account detection. *2017 International Conference on Computer Science and Engineering (UBMK)*. <https://doi.org/10.1109/ubmk.2017.8093420>

[10] Fletcher, E. (2022, January 25). *Social media a gold mine for scammers in 2021*. Federal Trade Commission. <https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-2021>

[11] Hackett, S. (2022, October 21). *How To Identify Fake Instagram Followers & Put A Stop To Them*. Kicksta Blog | Tips & Tricks to Get More Real Followers on Instagram. <https://blog.kicksta.co/how-to-identify-fake-instagram-followers-put-a-stop-to-them/>

[12] Henderi, H. (2021, March 1). *Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer | Henderi | International Journal of Informatics and Information Systems*. <https://ijiis.org/index.php/IJIIS/article/view/73>

[13] *How many accounts can I follow on Instagram?* (n.d.). Instagram Help. <https://help.instagram.com/408167069251249>

[14] *Introducing new authenticity measures on Instagram*. (2020, August 13). Instagram Blog. <https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram/>

[15] Kondeti, P., Yerramreddy, L. P., Pradhan, A., & Swain, G. (2020). Fake Account Detection Using Machine Learning. *Evolutionary Computing and Mobile Sustainable Networks*, 791–802. <https://doi.org/10.1007/978-981-15-5258-8_73>

[16] Lantz, B. (2019). *Machine Learning with R* (3rd ed.). Packt Publishing.

[17] Luis Torgo ltorgo@dcc.fc.up.pt. (n.d.). *SMOTE function - RDocumentation*. <https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE>

[18] Qiang Cao, Michael Sirivianos, Xiaowei Yang, & Tiago Pregueiro. (2012). Aiding the detection of fake accounts in large scale social online services. *Networked Systems Design and Implementation*, 15.

[19] Ramalingam, D., & Chinnaiah, V. (2018). Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers &Amp; Electrical Engineering*, *65*, 165–177. <https://doi.org/10.1016/j.compeleceng.2017.05.020>

[20] *Reducing inauthentic activity on Instagram*. (2018, November 19). Instagram Blog. <https://about.instagram.com/blog/announcements/reducing-inauthentic-activity-on-instagram>

[21] Roy, P. K., & Chahar, S. (2020). Fake Profile Detection on Social Networking Websites: A Comprehensive Review. *IEEE Transactions on Artificial Intelligence*, *1*(3), 271–285. <https://doi.org/10.1109/tai.2021.3064901>

[22] Sheikhi, S. (2020). An Efficient Method for Detection of Fake Accounts on the Instagram Platform. *Revue D’Intelligence Artificielle*, *34*(4), 429–436. <https://doi.org/10.18280/ria.340407>

[23] Siriseriwan, W. (2019). *A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE* (1.3.1) [Software]. <https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf>

[24] Van Der Walt, E., & Eloff, J. (2018). Using Machine Learning to Detect Fake Identities: Bots vs Humans. *IEEE Access*, *6*, 6540–6549. <https://doi.org/10.1109/access.2018.2796018>

1. [↑](#footnote-ref-1)