Why a bot is undetectable? An explainability-based study of misclassified automated accounts in social networks

Salvador Lopez-Joya[1,2][0009−0000−9290−6528], Jose A. Diaz-Garcia[1,2][0000−0002−9263−1402], M. Dolores Ruiz[1,2][0000−0003−1077−3173], and Maria J. Martin-Bautista[1,2][0000−0002−6973−477X]

[1] Research Centre for Information and Communications Technologies (CITIC-UGR), 18014 Granada

[2] Department of Computer Science and Artificial Intelligence, University of Granada, 18014 Granada

slopezjoya@ugr.es, {jagarcia, mdruiz, mbautis}@decsai.ugr.es

Abstract. Detecting bots on social media platforms is a major challenge, as these automated entities are constantly evolving to evade detection. In this study, we investigate the main features that contribute to the difficulty of bot detection. Leveraging the TwiBot-20 dataset, we analyze the characteristics of misclassified accounts and explore the reasons behind their erroneous classification. Our approach combines feature engineering, Machine Learning with Random Forest, and the interpretation of model predictions using SHAP (SHapley Additive exPlanations) values. We employ clustering techniques to identify patterns in feature contributions and provide insights into the complexities of distinguishing between human and automated accounts. Our findings highlight the nature of bot detection and the need for advanced methods to address the problem of social media manipulation.

Keywords: Bot detection · Social media · Machine learning · Explainability · SHAP values · Clustering · Data mining.

1 Introduction

The rise of social networks has transformed how people interact and engage online, creating ecosystems where automated programs, or bots, play a significant role in emulating human behaviour. While these bots can offer benefits, such as automating tasks for journalists or providing customer service [1], they also pose significant risks. The mass dissemination of political messages, fake news, and malicious links by bots has emerged as a potent tool for influencing and deceiving individuals at scale [2, 3].

Bots have become increasingly sophisticated, blurring the lines between human and automated activity. Although bot detection has received considerable attention from researchers in recent years [5], the challenge remains due to the evolving tactics used by bot creators to evade detection. The latter has served

as the motivation for this study. It has been observed that there is a need for information about the characteristics that make the task of detecting bots so complex. This article contributes to that need in the following ways:

- We provide an overview of the current state of the art in bot detection and its challenges, gaining detailed insight into one of the most widely used datasets in the bot detection literature using data science techniques.
- We provide valuable information and conclusions about the patterns found in the misclassified elements of this dataset to serve as a generalisation of the problems that can be found in bot detection.
- We use explainability techniques to provide a clear visual representation of the key features and how they contribute to bot detection, in order to gain knowledge about these accounts and improve future detectors.

The rest of the paper is organized as follows. In Section 2, we offer a comprehensive review of research that shares a similar approach to ours. Section 3 provides an intricate exploration of our methodology. The core of our investigation unfolds in Section 4, where we conduct a thorough analysis of the impact of various features on bot misclassification. Concluding remarks are presented in Section 5.

2 Related works

The understanding of what constitutes a bot in social media varies among researchers, leading to different definitions and classifications. Some define bots based on their level of automation, while others emphasize their similarity to human behaviour [9, 18]. Despite these variations, there is consensus that social media bots are accounts with a certain degree of automation, interacting within the social space [1, 9, 13, 14]. Understanding and defining these characteristics are crucial for developing effective bot detection mechanisms.

The literature distinguishes between three bot detection techniques: feature-based, graph-based and crowdsourcing techniques [8]. Among these techniques, feature-based methods are the most widely used. These methods leverage account metadata and user-generated content to identify bots. They can be categorized into account-based, content-based, and hybrid methods according to the area in which they are used [13].

Many datasets have been collected for bot classification. A public list of a large number of them can be found in the official Botometer repository [19]. Two of the most recent datasets are TwiBot-20 [7] and Twibot-22 [6]. These datasets have a variety of types of bots to detect, as well as the relationships of these accounts to other accounts, making it possible to implement graph-based methods. Despite the large number of bot detectors developed on these datasets, there is a significant percentage of accounts that are misclassified.

There are numerous works that perform feature engineering to enrich the classification of bots [4, 10, 16] but to our knowledge there are no articles that

study and evaluate which features influence the misclassification of these accounts, being this research the first approach for the profiling of undetectable bots in online social networks.

3 Our approach

The problem we face is a classification problem where we have a set of users U, denoted $U = \{u_1, u_2, u_3, ..., u_n\}$, and a set of tweets T for each user denoted $T_u = \{t_1, t_2, t_3, ..., t_m\}$. The goal is to make a prediction with binary variables, where $y(u_i) = 0$ implies that an account is real and $y(u_i) = 1$ implies that the account is a bot, where $i \in \{1, \ldots, n\}$. We denote the set of true labels as Y, where y_i is the true label for the i-th instance

$$Y = \{y_1, y_2, \ldots, y_n\} \tag{1}$$

and the set of predictions as $\hat{Y}$, where $\hat{y}_i$ represents the predicted label for the i-th instance

$$\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\} \tag{2}$$

The final aim of this paper is to study the features and behaviour of the set where predictions and true labels differ. We will denote this set as E and define it as follows:

$$E = \{u_i \mid \hat{y}_i \neq y_i\} \tag{3}$$

3.1 Methodology

The dataset chosen for this experiment is the Twibot-20 [7], this decision is justified because it is one of the most recent datasets, along with the Twibot-22 [6], and unlike the latter, many more studies have used the Twibot-20 to date, thus providing us with more comparative results. Furthermore, despite the number of bot detectors tested in this dataset, the maximum accuracy obtained is around 87% [6, 11, 12], making it an excellent candidate for studying why there are still bots that we can not classify.

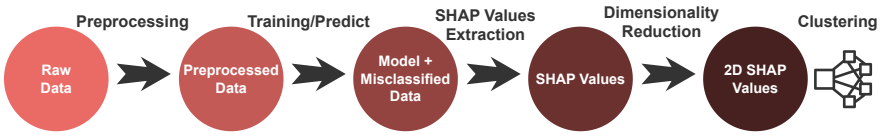


Fig. 1: Methodology flow diagram

The step-by-step process we have followed can be seen in Fig. 1. We initiate by preprocessing the dataset, involving data cleaning, categorical variable encoding, handling missing values, and numerical feature scaling. In order to enrich the

dataset and improve the results, we have selected some additional features from other works that have been shown to have some influence on the performance of their respective models [19]. These attributes are inferred from the information we already have, and we classify them as follows:

– Features inferred from the user account:

$$user\_age = \frac{(reference\_date - timestamp).dt.total\_seconds()}{86400} \tag{4}$$

$$tweet\_freq = \frac{statuses\_count}{user\_age} \tag{5}$$

$$followers\_growth\_rate = \frac{followers\_count}{user\_age} \tag{6}$$

$$friends\_growth\_rate = \frac{friends\_count}{user\_age} \tag{7}$$

$$favourites\_growth\_rate = \frac{favourites\_count}{user\_age} \tag{8}$$

$$listed\_growth\_rate = \frac{listed\_count}{user\_age} \tag{9}$$

$$followers\_friends\_ratio = \frac{followers\_count}{friends\_count} \tag{10}$$

– Features inferred from the tweets:

$$reply\_count\_mean = \frac{\sum_{i=1}^{n} reply\_count_i}{n} \tag{11}$$

$$num\_hashtags\_mean = \frac{\sum_{i=1}^{n} num\_hashtags_i}{n} \tag{12}$$

$$num\_urls\_mean = \frac{\sum_{i=1}^{n} num\_urls_i}{n} \tag{13}$$

$$num\_mentions\_mean = \frac{\sum_{i=1}^{n} num\_mentions_i}{n} \tag{14}$$

For the classification task we chose Random Forest. Random Forest is a machine learning algorithm widely employed for bot detection in social media platforms due to its effectiveness. It reduces overfitting by combining predictions from multiple decision trees and is less susceptible to noise and outliers present in the data. But there is another reason why it has been chosen: its explainability. Random Forest provides the predictive value of each feature and the final decision tree can be visualised.

Furthermore, we employ SHAP (SHapley Additive exPlanations) [15] values to provide insights into the key features influencing the model's predictions. SHAP values offer a comprehensive explanation of individual predictions by computing the contribution of each feature.

Finally, clustering was also performed on these SHAP values to identify patterns in the features. This approach aimed to facilitate a deeper understanding of their interactions and impact on the model's predictions. The clustering algorithm used is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) which is particularly advantageous in scenarios where the data exhibit varying densities and irregular shapes of clusters. In order to perform this clustering, a dimensionality reduction technique has been applied, specifically UMAP (Uniform Manifold Approximation and Projection) [17]. This technique was chosen over PCA (Principal Component Analysis) due to the high dimensionality of the dataset and the existence of non-linear relationships between its features.

4 Results and discussion

In this section, we will visualise the results of our analysis. We will present the results of the model, the general SHAP plot, the clustering performed and some SHAP plots of the most representative elements of each cluster.

The results of the Random Forest model, using a cross-validation with 10 partitions, gave us a mean accuracy of 0.8179 and the confusion matrix of this model can be seen in Table 1. As we can see, the model is significantly more likely to be wrong in its prediction of a human when the account is a bot.

		Predicted Class		
		Negative	Positive	Total
True Class	Negative	378	165	543
	Positive	50	590	640
	Total	428	755	1183

Table 1: Confusion Matrix of TwiBot-20 using a Random Forest.

From this model, a general SHAP plot has been made to identify which features are the most important and how they affect the classification. This graph

can be seen in Fig. 2. In the graph, it can be observed that account verification is a good way to determine whether an account is real, but we can also see that a high average number of URLs in tweets contributes to an account being considered a bot, the same is true for the number of hashtags, the number of mentions and the number of friends. The opposite is true for the average number of emoticons per tweet. Looking at features like *followers_friends_ratio*, *followers_count*, *listed_count*, *followers_growth_rate* and *listed_growth_rate*, we can see how high values contribute to the prediction of a human account but, when these values are low, the contribution to the prediction can be a bit more ambiguous. Another interesting feature is *geo_enabled*, looking at this feature we can see that when it is true it contributes to the prediction of a human account, on the other hand when it is false it contributes to the prediction of an automated account.

Fig. 2: Training data general SHAP plot.

As can be seen in the confusion matrix in Table 1, there are 215 misclassified elements. In order to get information about the reason for this misclassification, these elements were extracted and the SHAP values of each of them were obtained. By applying UMAP, a clustering was carried out, which can be seen in Fig. 3.

The estimated number of clusters using DBSCAN is 5 and these clusters cover the 86% of the misclassified data. From each cluster, some representative elements have been selected and their SHAP plots have been visualised to obtain information on the contribution of each feature in the prediction. We selected

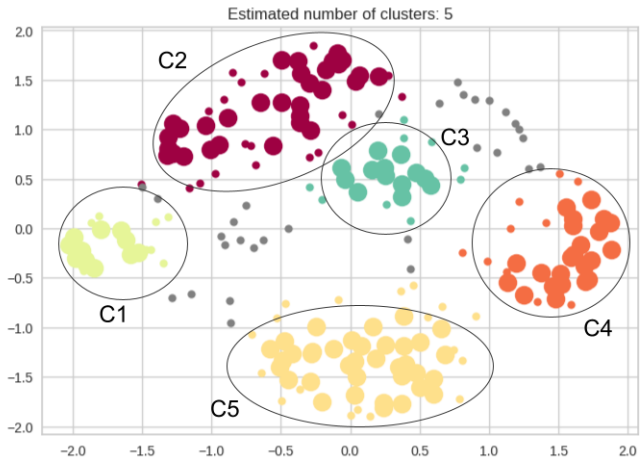
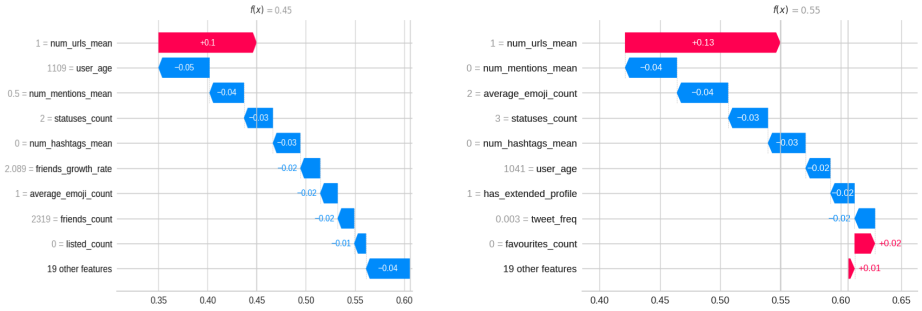


Fig. 3: Clustering of SHAP values applying dimensionality reduction.

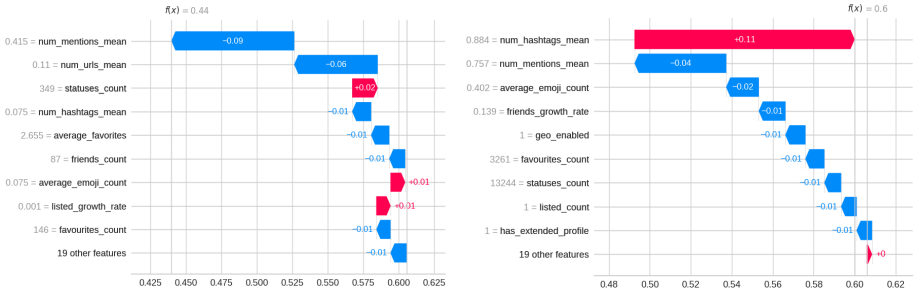
these elements from those closest to the centroid of each cluster and, by visualising them, we verified that they were similar to their neighbours and therefore representative of the cluster. These plots can be seen in Fig. 4 and Fig. 5. The clusters found are the following:

Cluster 1. Looking at the elements of this cluster in Fig. 4, we can see a pattern: the average number of URLs per tweet contributed to the bot class prediction. In the left graph we can see an incorrect prediction of an account that is actually a bot. In this case, the average number of URLs contributed to the correct prediction, but not enough, while all other features contributed to the incorrect prediction. The graph on the right shows an incorrect prediction of a human account. In this case, the average number of URLs was the cause of the model's failure. Most of the other features have contributed correctly to the prediction. Looking more closely at these two accounts, we realise that the number of tweets is extremely low (i.e., 2 tweets in the first account and 3 in the second), if there are not enough tweets some metrics may not be informative, in fact, they may cause the model to fail.

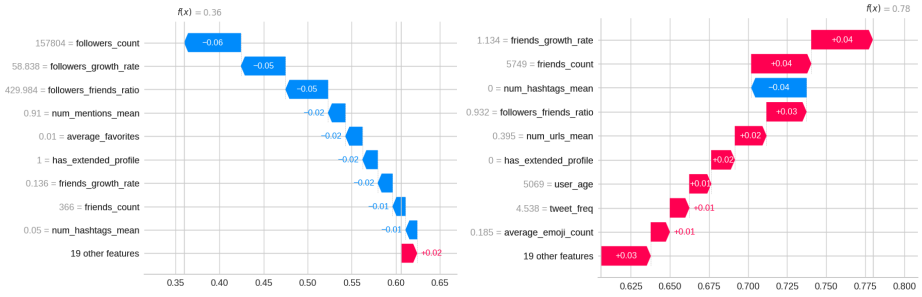
Cluster 2. The common element between the members of this cluster, as we can see in Fig. 4, is the contribution of the average number of mentions per tweet to the prediction of the human class. As in Cluster 1, we can see an incorrect prediction of a real account in the left part and an incorrect prediction of an automated account in the right part. The graph showing the incorrect prediction of a human shows that both the average number of mentions and the average number of URLs were decisive features for the failure of the model, whereas in the misclassification of the bot, the average number of mentions contributed to a



Cluster 1



Cluster 2

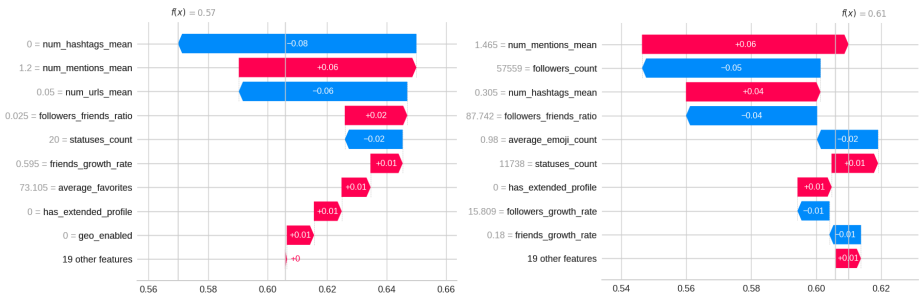


Cluster 3

Fig. 4: SHAP values of some representative elements from clusters 1, 2 and 3. Left figures represent Human incorrect predictions. Right figures represent bot incorrect predictions.

good prediction, which ended up being wrong because of the average number of hashtags. Looking more closely at both examples, we can see that the left graph is about a bot that covers sports events. This bot posts mostly plain text with no URLs and moderate mentions. The right graph is a real account promoting an event using a hashtag. This information suggests that while certain patterns are used more often by bots or human accounts, their absence does not necessarily indicate the opposite. A bot may post many tweets without using URLs and mentions, and a real account may use hashtags to promote its event.

Cluster 3. This cluster (see Fig. 4) does not have such a clear pattern among its elements as the previous ones, but if we examine its elements we can see that the features related to followers and friends tend to be the ones that have contributed the most to the prediction. Following the same procedure we can see that in both graphs most of the features contributed to the wrong prediction. Looking at these cases in more detail, we realise that there are accounts where it is difficult to determine a label. There are accounts for memes, political opinions, influencers and others that could have been automated accounts at some point.



Cluster 4 Cluster 5

Fig. 5: SHAP values of some representative elements for bot incorrect prediction

Clusters 4 and 5. Finally, Fig. 5 shows one element from Cluster 4 and one from Cluster 5. These clusters do not contain bots and are composed only of misclassified examples of real accounts. In Cluster 4, we can see how the average number of mentions and the average number of hashtags per tweet contribute to the model assigning the bot label to the account. This is true for almost all the elements in this cluster. Inspecting this account further, we can see that it is a journalist and activist who uses mentions and hashtags very frequently. In Cluster 5, we can see how the average number of hashtags and the average number of URLs condition the model, this time correctly, towards the prediction of being a bot. In this example we encounter the same problem as with the elements in Cluster 1, i.e. the account does not contain enough tweets.

From the information obtained about these patterns, we can draw some conclusions:

1. Many of the metrics based on user content are misleading if the number of posts by that user is insufficient. One possible way to solve this problem is to set a threshold of a minimum number of posts when evaluating accounts and to take other types of posts into account. This could be done by setting a minimum number of posts when evaluating accounts and taking other types of characteristics into account.
2. While the number of URLs, mentions and hashtags are often anomalous characteristics of automated accounts, they are not sufficient evidence that the account is a bot. There are accounts that make unusual use of these tools and that does not necessarily mean they are bots.
3. There are accounts that are really complicated to classify, even for a human, and there are also outlier accounts such as streamers, influencers, actors, etc. Including more accounts of this type could be an interesting option for the model to consider.

5 Conclusion

The study presented in this paper offers insights into the challenges of detecting bots on social media platforms. By analyzing misclassified accounts, we have identified key features that contribute to the difficulty of bot detection. Our approach combines feature engineering, machine learning using Random Forest, and the interpretation of model predictions using SHAP values. Through clustering techniques, we have uncovered patterns in feature contributions in order to make the task of distinguishing between humans and bots less complex.

This analysis has focused primarily on 𝕏, but it's important to note that other platforms such as Instagram, Reddit or LinkedIn have received less research attention and could also could also be a good avenue for future research.

Acknowledgments

The research reported in this paper was supported by the DesinfoScan project: Grant TED2021-129402B-C21 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and FederaMed project: Grant PID2021-123960OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309).

References

1. Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., Alomari, D.M.: Machine learning-based social media bot detection: A comprehensive literature review. Social Network Analysis and Mining **13**(1), 20 (2023)

2. Beisecker, S., Schlereth, C., Hein, S.: Shades of fake news: How fallacies influence consumers' perception. European Journal of Information Systems **33**(1), 41–60 (2024)
3. Bovet, A., Makse, H.A.: Influence of fake news in twitter during the 2016 us presidential election. Nature communications **10**(1), 7 (2019)
4. Cardaioli, M., Conti, M., Di Sorbo, A., Fabrizio, E., Laudanna, S., Visaggio, C.A.: It'sa matter of style: Detecting social bots through writing style consistency. In: 2021 International Conference on Computer Communications and Networks (ICCCN). pp. 1–9. IEEE (2021)
5. Cresci, S.: A decade of social bot detection. Communications of the ACM **63**(10), 72–83 (2020)
6. Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., et al.: Twibot-22: Towards graph-based twitter bot detection. Advances in Neural Information Processing Systems **35**, 35254–35269 (2022)
7. Feng, S., Wan, H., Wang, N., Li, J., Luo, M.: Twibot-20: A comprehensive twitter bot detection benchmark. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 4485–4494 (2021)
8. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM **59**(7), 96–104 (2016)
9. Gorwa, R., Guilbeault, D.: Unpacking the social media bot: A typology to guide research and policy. Policy & Internet **12**(2), 225–248 (2020)
10. Ilias, L., Roussaki, I.: Detecting malicious activity in twitter using deep learning techniques. Applied Soft Computing **107**, 107360 (2021)
11. Liang, L., Wang, X., Liu, G.: Semantic information mining and fusion method for bot detection. In: International Conference on Artificial Neural Networks. pp. 270–282. Springer (2023)
12. Liu, F., Li, Z., Yang, C., Gong, D., Lu, H., Liu, F.: Segcn: a subgraph encoding based graph convolutional network model for social bot detection. Scientific Reports **14**(1), 4122 (2024)
13. Lopez-Joya, S., Diaz-Garcia, J.A., Ruiz, M.D., Martin-Bautista, M.J.: Bot detection in twitter: An overview. In: International Conference on Flexible Query Answering Systems. pp. 131–144. Springer (2023)
14. Loyola-González, O., Monroy, R., Rodríguez, J., López-Cuevas, A., Mata-Sánchez, J.I.: Contrast pattern-based classification for bot detection on twitter. IEEE Access **7**, 45800–45817 (2019)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
16. Mbona, I., Eloff, J.H.: Feature selection using benford's law to support detection of malicious social media bots. Information Sciences **582**, 369–381 (2022)
17. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arxiv 2018. arXiv preprint arXiv:1802.03426 (1802)
18. Stieglitz, S., Brachten, F., Ross, B., Jung, A.K.: Do social bots dream of electric sheep? a categorisation of social media bot accounts. arXiv preprint arXiv:1710.04044 (2017)
19. Yang, K.C., Varol, O., Hui, P.M., Menczer, F.: Scalable and generalizable social bot detection through data selection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 1096–1103 (2020)