Unmasking Fake Social Network Accounts with Explainable Intelligence

Eman Alnagi[1], Ashraf Ahmad[2], Qasem Abu Al-Haija*[3], Abdullah Aref[4]
Department of Computer Science, King Hussein School of Computing Sciences,
Prince Sumaya University for Technology, PO Box 1438, Amman 11941, Jordan[1,2,4]
Department of Cybersecurity-Faculty of Computer & Information Technology,
Jordan University of Science and Technology, PO Box 3030, Irbid 22110, Jordan[3]

***Abstract*—The recent global social network platforms have intertwined a web connecting people universally, encouraging unprecedented social interactions and information exchange. However, this digital connectivity has also spawned the growth of fake social media accounts used for mass spamming and targeted attacks on certain accounts or sites. In response, carefully constructed artificial intelligence (AI) models have been used across numerous digital domains as a defense against these dishonest accounts. However, clear articulation and validation are required to integrate these AI models into security and commerce. This study navigates this crucial turning point by using Explainable AI's SHAP technique to explain the results of an XGBoost model painstakingly trained on a pair of datasets collected from Instagram and Twitter. These outcomes are painstakingly inspected, assessed, and benchmarked against traditional feature selection techniques using SHAP. This analysis comes to a head in a demonstrative discourse demonstrating SHAP's suitability as a reliable explainable AI (XAI) for this crucial goal.**

***Keywords*—Explainable Artificial Intelligence (XAI); Shapley Additive exPlanations (SHAP); feature selection; fake accounts detection; social media**

I. Introduction

With the rapid development of the Internet, social networks have become widespread platforms that connect people worldwide to socialize, communicate, and share knowledge. Different types of social networks have invaded the Internet. Some are used for social activities and connections, such as Facebook and Twitter. Others, such as YouTube, Instagram, and Pinterest, are used for sharing videos and pictures. Some are used to build professional connections, such as LinkedIn, and others to create science and research networks, such as ResearchGate. All these social networks with public data scattered over the Internet urged several malicious parties to take advantage of this situation. Fake accounts have made it easy for such parties to reach naive people's accounts for spreading spam messages, blackmailing, or hacking. The increasing number of fake accounts all over social networks has increased the necessity of detecting them. Artificial intelligence (AI), in general, and machine learning (ML) algorithms, in particular, are some common approaches in the literature used to detect whether an account is fake. These types of prediction algorithms have succeeded in the detection of fake accounts. Nevertheless, in most cases, the accuracy of these algorithms is less than 100%. False positive and negative results keep raising and decreasing consumers' trust in these models, making AI a black box that needs to be explained to convince consumers of its importance and usability. Explainable AI (XAI) techniques [1] are algorithms proposed to explain the results of this black box. AI programmers have proposed and programmed various approaches to explain the outcomes of their models. Some of them work on tabular data, others on images or text. In this paper, Shapley Additive exPlanations (SHAP) [2] has been selected as one of these explainable AI algorithms, which can be used on tabular data. This algorithm was selected to explain a trained model on two datasets prepared for the fake account detection task. The two datasets have been trained using XGBoost [3].

This work analyzes the results of the SHAP algorithm and compares them with the traditional feature selection algorithms that highlight the important features of a dataset. This paper is organized as follows: Section II provides the related work and reviews some state-of-the-art work in the same study area. Section III discusses the methodology and details the system development phases. Section IV, the result and discussion, illustrates the results, discussion, and analysis. Finally, section V concludes the work discussed in this research and illustrates the limitations and future work.

II. Related Research

Detecting fake accounts in social networks is a common task tackled in the literature, using different classification algorithms and datasets from different platforms. Authors of [4]–[11] have worked on datasets collected from Twitter to detect fake or bot accounts. Many account features have been gathered in these datasets. Some of them are related to the profile itself, such as the number of followers, the number of following statuses, whether the account is protected or verified, including a profile picture, and many others. Other features related to the tweet content include the number of URLs in a tweet, mentions, hashtags, emojis, etc. Authors [12] and [13] have worked on Facebook datasets. Examples of the features they used for prediction are the existence of a bio, a workplace, family members, check-ins, the number of friends, the number of followers, and many others. In the same context, in [9], the authors have also tackled datasets of LinkedIn, where certain features such as the number of languages, number of connections, number of skills, and others were collected. Moreover, several AI algorithms have been used to accomplish this task. The authors of [4, 6-9, 11] have used several machine learning algorithms, such as XGBoost (XGB), Random Forest (RF), Support Vector Machine (SVM), AdaBoost, and Logistic Regression (LR). Others used Naive Bayes (NB) [14] and K-Nearest Neighbors (KNN) [12]. A

TABLE I. FEATURE DESCRIPTION OF INSTAGRAM DATASET

id	Feature	Description
1	Profile_pic	A boolean feature indicates whether the account has a profile picture or not
2	nums/length_username	The ratio of the number of digits in a user name and the length of the username
3	fullname_words	Number of words in the account full name
4	nums/length_fullname	The ratio of the number of digits in an account's full name and the length of the full name
5	name==username	A boolean feature that indicates whether the full name is similar to the username
6	description_length	The length of the account bio
7	external_URL	A boolean feature that indicates whether the account has an external URL or not
8	private	A boolean feature that indicates whether an account is private or not
9	#posts	Number of posts published by the account
10	#followers	Number of followers
11	#follows	Number of accounts this user is following
12	**fake**	**The binary class indicates whether the account is fake or not**

TABLE II. FEATURE DESCRIPTION OF TWITTER DATASET

id	Feature	Description
1	screen_name_length	Number of characters in a screen_name
2	location	A boolean feature that indicates whether a location is specified or not
3	has_description	A boolean feature that indicates whether the account includes a description or not
4	followers_count	Number of followers
5	friends_count	Number of friends
6	listed_count	Number of listed accounts
7	favourites_count	Number of favourites
8	verified	A boolean feature that indicates whether an account is verified or not
9	statuses_count	Number of statuses in the account
10	default_profile	A boolean feature that indicates whether the account uses the default profile or not
11	default_profile_image	A boolean feature that indicates whether the account has an extended profile or not
12	has_extended_profile	A boolean feature that indicates whether the account has an extended profile or not
13	name_length	Number of characters in the username
14	**bot**	**The binary class indicates whether the account is a bot or not**

survey published in 2021 has been conducted on cybersecurity AI [15], reviewing many AI algorithms that have been applied to many scopes of cybersecurity, such as intrusion detection, spam detection, phishing, and fake news detection. They have added a small section about XAI in cybersecurity, indicating that this field still needs more research on XAI. The researchers in [16] also looked at two XAI methods, LIME and Saliency Map, and compared them to explain a trained model for website fingerprinting attacks. The most related work to this paper is the work of [17], where Twitter bot detection has been applied and explained using the LIME XAI approach.

III. PROPOSED XAI DETECTION SYSTEM

This section provides details about the main components of the proposed XAI detection system. We describe the two datasets (Dataset 1: Instagram + Dataset 2: Twitter). We describe the feature selection process implemented in this paper. Then, we'd like to present the learning techniques for developing the classification model. After that, we describe, using XAI, the SHAP model to explain the proposed classification dynamics. Finally, we define our experimental setup environment.

1) Datasets selection: For this research, a meticulous selection process has been undertaken to identify two distinct datasets, each sourced from prominent social media platforms (Instagram and Twitter). These datasets are representative reservoirs of the unique dynamics and user behaviors exhibited on these platforms, enriching our analysis's depth and breadth.

- Instagram Dataset: This dataset consists of data about Instagram accounts. It is a public dataset published on the Kaggle website [18]. Each instance is labeled as either fake or not. It consists of 676 records and 11 features. Data pre-processing steps are optional for this dataset before it has been used in the training. Table I displays a description of these features.
- Twitter Dataset: This dataset has been created to detect whether a Twitter account is a bot. Among other datasets by [19], it has been proposed to study social spam bots. The dataset consists of 2797 instances and 19 features, but some have been removed, such as "id" and "id_str". The selected ML model has modified others to be more suitable for training. For example, the original dataset contains the screen name as a text; this has been changed to become the length of the screen name. The location exists as the name of a place; it has been changed to become a Boolean feature that indicates whether the location is specified or not. This resulted in 13 features in the dataset. Table II describes the final set of features.

A. Feature Selection

Feature selection methods are usually used in classification tasks to reduce the dimensionality of large datasets [20]. Dimensionality reduction affects the performance of classification models since such models are trained on a subset of dataset features and thus save computational time. In this work, a feature selection approach based on a Random Forest classifier is used to highlight the essential features in the datasets and compare them with the results of the XAI algorithm.

B. Classification-based XGBoost Model

XGBoost has been selected as the classification model to be explained. XGBoost is an ensemble of decision trees with gradient tree boosting [3]. It has been selected as the primary classifier since it has been widely used in literature to predict fake social media accounts. Fig. 1 shows how XGBoost works. XGBoost works by joining several weak learners (decision trees), each trained on a subset of the dataset to establish a strong learner. The stronger learner tends to be highly efficient, flexible, and portable.

C. XAI-based SHAP model

The main contribution of this research is to explain the results of the classification model trained on both datasets with an Explainable AI approach. SHAP has been selected for this purpose. According to Rothman [22], SHAP's intuition has been raised from game theory. In game theory, each player

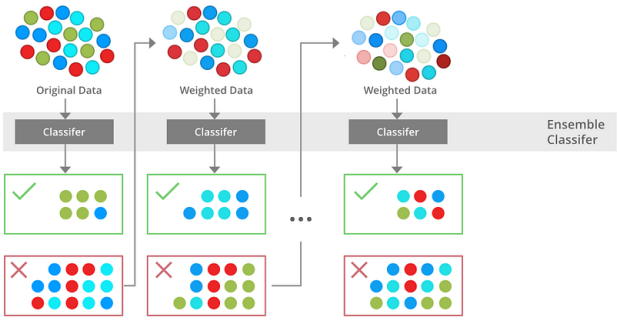

Fig. 1. How XGBoost works [21].

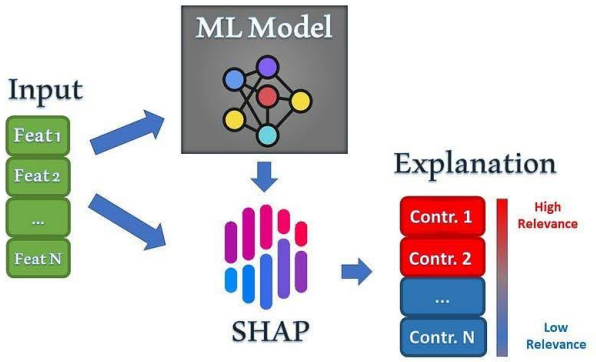

Fig. 2. How SHAP works [23].

has a contribution to a game that yields the final result. So, SHAP has been created to approximate the contribution of each feature in a dataset to predict a correct or wrong class. Using SHAP, several charts can be plotted to reveal the secrets of the black box of AI. Some of these charts describe the global effect of the features as an approximation of this effect on all instances in the datasets. Other charts concentrate on a single instance or a range of instances. Several types of these charts have been used to explain the results of training the two datasets. Fig. 2 below shows how the SHAP model works. SHAP is a Model-agnostic, post-hoc method that takes several input features concurrently trained using the ML model(s) to explain/interpret (level of relevance) for the feature attribution and presents the level of model trustworthiness.

D. Experiments Environment

Google Colaboratory [24] has been used as the programming platform for the Python programming language. Python libraries have been used to apply feature selection tasks as the first step. Then, the Special Python Library, SHAP, was used to apply the explainer function and produce descriptive charts to help analyze the resulting prediction from the classification model.

IV. Experimental Results and Discussion

This section is dedicated to the thorough presentation and intuitive assessment of our experimental results, which are

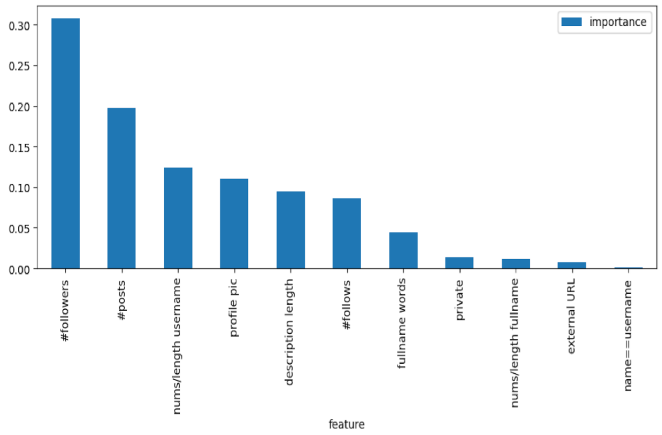

Fig. 3. The feature selection method results on the instagram dataset.

TABLE III. Confusion Matrix of Instagram Dataset

Confusion Matrix (Instagram Dataset)		Actual	
		Fake	Not Fake
Predicted	Fake	59	4
	Not Fake	8	69
Support		67	73

dissected and discussed for each dataset in isolation (the Instagram and Twitter datasets). Table III shows consfusion matrix of Instagram dataset.

A. Experimental Results for Instagram Dataset

A feature selection method was applied to the dataset as a first step to extract the important features that a classification model can rely on. Nevertheless, the classification model is then trained on the complete feature list so that the XAI algorithm can highlight and explain the effect of these features on the classification results. The feature selection step is applied only for comparison purposes. According to the feature selection method, it has been found that the top five features in this dataset are #followers, #posts, username length, the existence of a profile picture, and description length, as illustrated in Fig. 3. This algorithm only highlights the features most correlated to the class label to be predicted. Nevertheless, it cannot be decided from these results how these features affect the classification results, raising the need for the XAI algorithm.

Training the dataset with the XGBoost classifier yielded an accuracy of 91%. Table IV illustrates the results of other evaluation metrics for this experiment: precision, recall, and f1-score. When applying the SHAP explainer, results have been illustrated as charts. Fig. 4 displays the global bar plot chart, which explains how each feature has affected the prediction results in all instances in the dataset. Fig. 4 demonstrates that the presence of a profile picture is the second most effective and significant factor in the prediction, after the number of followers. Other features, such as the similarity between the name and the username and the existence of an external URL, have less effect on the prediction.

It can be noticed that this chart considers the same top five features extracted by the feature selection method in Fig. 3,

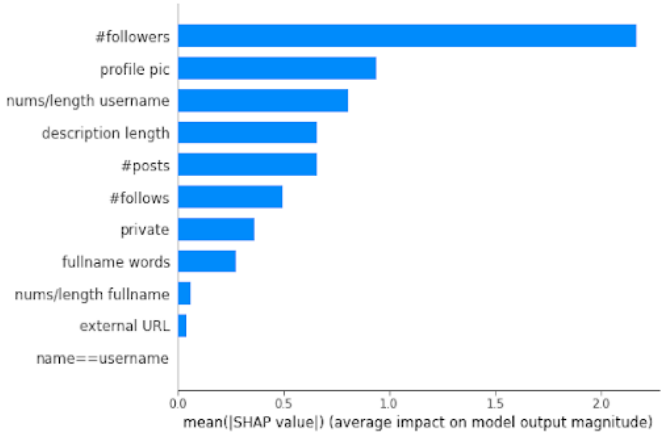

Fig. 4. Global bar plot chart: Global effect of features of the instagram dataset on prediction.

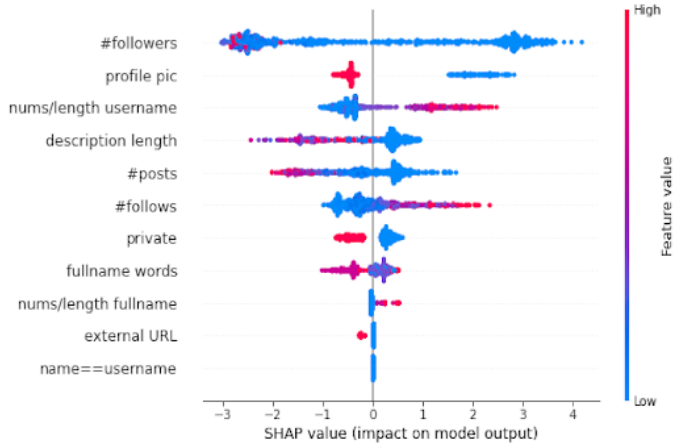

Fig. 5. SHAP Summary plot chart: Global effect of features of the instagram dataset on prediction.

TABLE IV. RESULTS OF INSTAGRAM DATASET

Class	Precision	Recall	F1-score
Fake	0.94	0.87	0.9
Not Fake	0.88	0.95	0.91

but with changing their order. Nevertheless, this effect may be considered positive or negative. Sometimes, a feature may lead the model to predict wrongfully. More detailed charts may explain this effect. Another chart that illustrates the global importance of features but with additional information is the summary plot in Fig. 5. Each feature is represented in this chart to illustrate its importance from the highest to the lowest. Other information is added using the colors that represents the feature value. For example, it can be noticed how the blue color in the #followers feature indicates that the low number of followers leads to a higher SHAP value, which explains how a low number of followers can affect the prediction of a fake account (value =1). The red dots in the same feature indicate a high number of followers, and they are concentrated on the negative side of the SHAP values, which represent the class value (0), not the fake account. Nevertheless, blue dots (low values) on the negative side might mislead the prediction, as illustrated in the coming charts. Another example is #posts; most of the blue dots reside in the positive SHAP value, which indicates that when the number of posts is low, the account is more likely to be fake. The feature #follows, however, shows the opposite behavior. It can be noticed from the red dots concentrated on the positive side of SHAP values that when the account follows a high number of other accounts, it is more likely to be a fake one. This result can be logically explained, especially for fake spam accounts; their behavior tends to follow as many accounts as possible to spread spam advertisements or news.

For a deeper look at the importance of features, a local bar plot chart illustrates the effect of features on a specific instance. Fig. 6 illustrates the results of four instances. It can be noticed from Fig. 6(a) and 6(b) that the model has succeeded in predicting the correct class. In both cases, the number of followers has been the most critical feature of this success. In Fig. 6(a), the model considered the account fake since the number of followers is low (730), considering that the maximum number of followers in the dataset is 15,338,538. As shown in Fig. 6(b), the number of followers is still low compared to the maximum number in the dataset. Nevertheless, the model succeeded in predicting that the account was not fake. Other features, such as the number of posts or the existence of a profile picture, may have participated in this prediction. As for Fig. 6(c) and 6(d), the model must include the correct prediction in both cases. In Fig. 6(c), the model misclassified the account as fake. In this case, a logical reason may be the number of followers alone. However, when considering Fig. 6(d), it will be noticed that this feature, #followers, has misled the prediction since although the account is a fake one, the model predicted it to be not fake because of the large number of followers that reached 19,512. Table V shows the confusion matrix of Twitter dataset.

TABLE V. CONFUSION MATRIX OF TWITTER DATASET

Confusion Matrix (Twitter Dataset)		Actual	
		Bot	Not Bot
Predicted	Bot	232	23
	Not Bot	38	267
Support		270	290

TABLE VI. RESULTS OF TWITTER DATASET

Class	Precision	Recall	F1-score
Bot	0.91	0.86	0.89
Not Bot	0.88	0.92	0.9

B. Experimental results for Twitter Dataset

The same steps have been applied to Dataset2. Fig. 7 illustrates the results of the feature selection method, which considered the top five features: friends count, favorites count, followers count, statuses count, and whether the account is verified. Then, the dataset was trained using an XGB classifier to yield a 90% accuracy; other evaluation metrics results are displayed in Table VI. Finally, the SHAP explainer was applied, which resulted in several charts.

Fig. 8 illustrates that the number of friends is the most

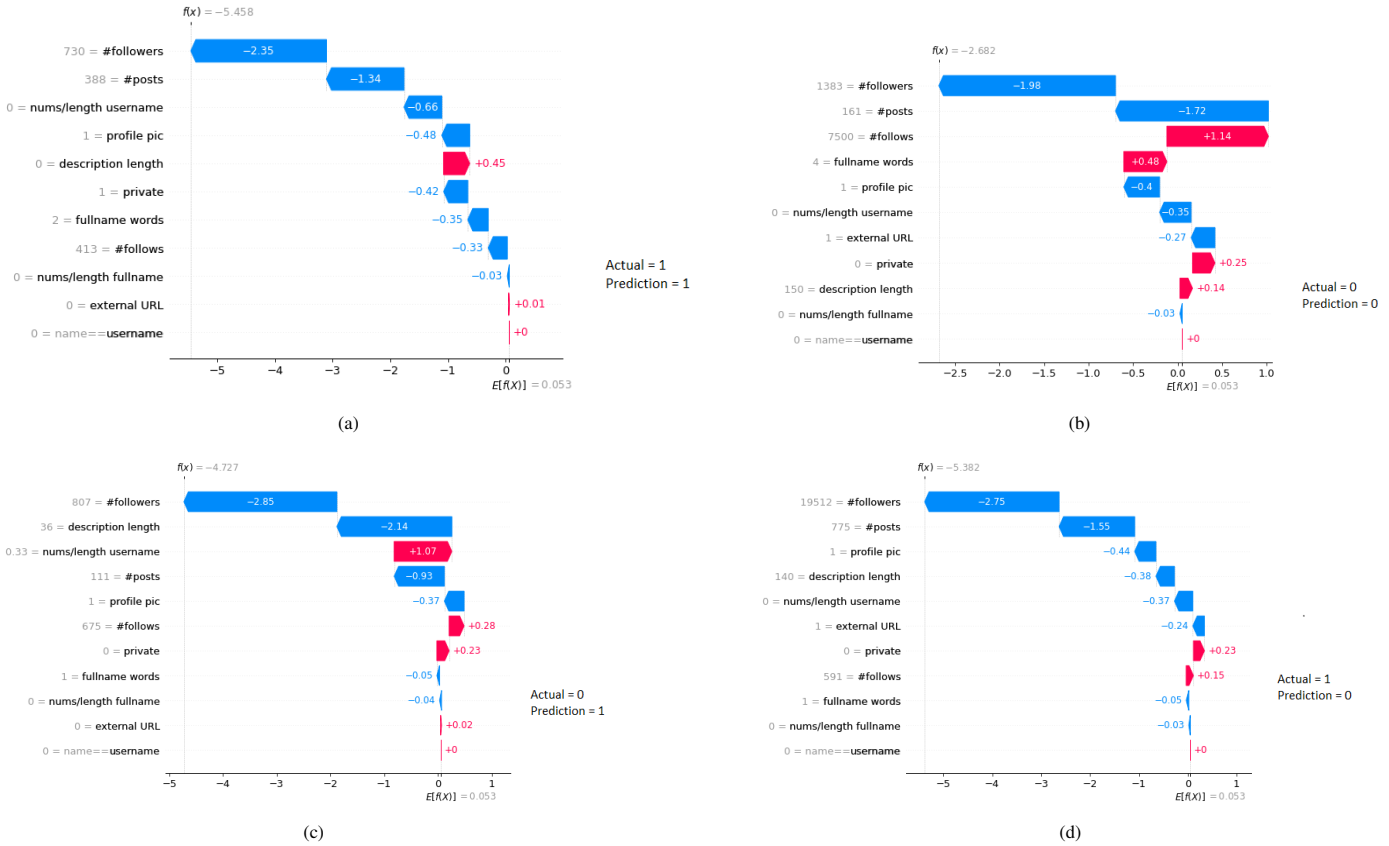

Fig. 6. Local bar plot charts - Dataset1.

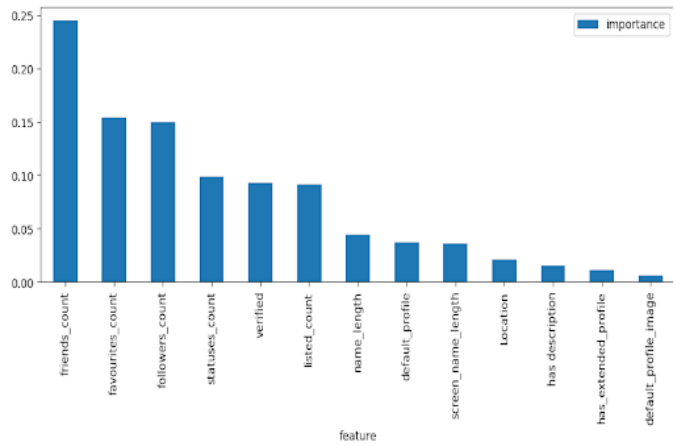

Fig. 7. The feature selection method results on the twitter dataset.

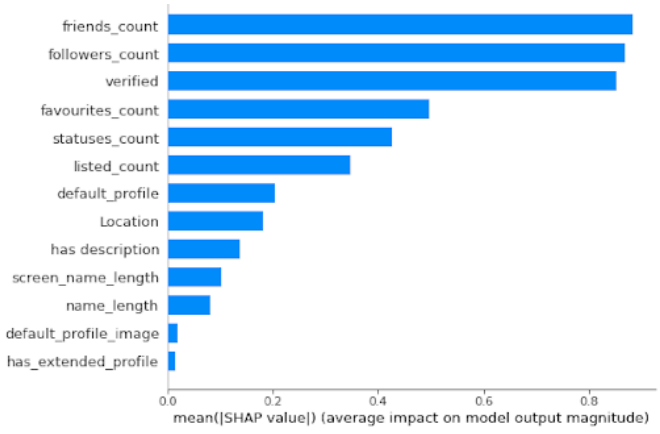

Fig. 8. Global bar plot chart: Global effect of features of the twitter dataset on prediction.

important feature in the dataset. The number of followers comes in the second rank and whether the account is verified. The least important features are whether the account has an extended profile and uses the default profile image. Compared with the results in Fig. 7, these are considered similar, but with the change of ordering the top five features. Surely, these findings are realistic since a bot account tends to follow as many accounts as it can; on the contrary, very few accounts follow a bot account since most people tend to follow accounts with either familiar users or at least accounts that contain valuable content, which is not usually the case in bot accounts. The same applies to friends count since even if the bot account tries to send friendship requests to other accounts, the users in these accounts have the right to decide whether to accept this friendship request or not, and usually, people do not add friends they do not know or at least have mutual friends with them.

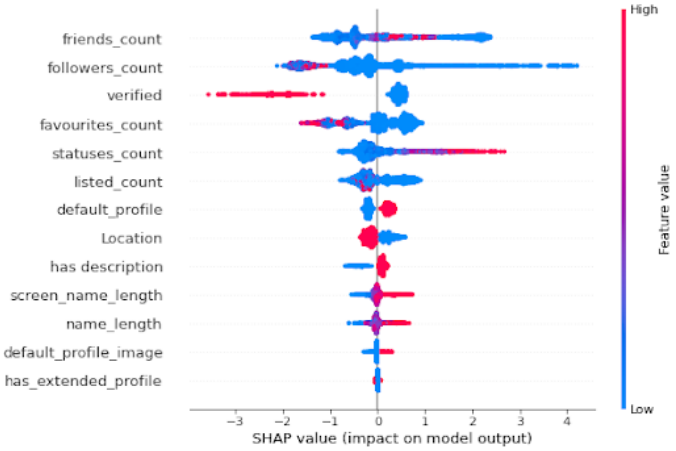

Fig. 9. SHAP Summary plot chart: Global effect of features of the twitter dataset on prediction.

Fig. 9 shows some logical influence of features and some strange ones. As for the number of followers, it can be noticed that the red dots (high values) tend to take the prediction to the negative SHAP values, which are the predictions of a bot account. Also, the verified feature is surely predicted not to be a bot when it is high (value = 1). Nevertheless, the number of friends is strange since most red dots reside in the positive SHAP values, which means the model might be misled to predict false positive bots. Since, logically, bots should not have a large number of friends.

Fig. 10 displays another type of chart, the Decision Plot, to explain the Local Bar Plot chart. Fig. 10(a) shows that the illustrated instance has been predicted to be a bit, even though it is not. This chart illustrates how the number of friends, which is 5, led to this prediction. Also, when comparing this result with Fig. 10(b), this will show how the decision of the final prediction has passed through the features to reach the false positive prediction. It can be noticed that the verified and the number of favorites may share the responsibility for this decision. Fig. 10(c) illustrates another instance falsely predicted to be not a bot, although it is a bot. Again, the number of friends is the dominant feature in controlling this decision, which is not logical because of its small value (153) compared with the maximum number of friends in the dataset, which is 2,056,668.

V. Conclusion and Future Work

An Explainable AI approach, SHAP, has been used in this paper to explain the results of fake Instagram accounts and Twitter bot detection. The detection task was applied using the XGBoost classifier, and the results were explained using SHAP. The feature selection method is used to verify the XAI algorithm's selection of highly effective features. Then, a global feature importance analysis and a local feature importance analysis of certain instances were conducted. SHAP has been proven to be a proper XAI approach for this task since it highlighted the most important features that affected the ML algorithm and directed it to the final prediction, resulting in high performance with low rates of false negatives and false positives predictions. Our work has some limitations, though; Only two social network platforms have been studied (Instagram and Twitter), and the work can be extended to include datasets of Facebook, Telegram, and other common platforms. Also, other types of XAI approaches should be analyzed in this work. Additional and deeper analysis of the dependency between features will be studied in future work. Also, other XAI approaches, such as LIME and EAI5 [25], will be used to explain fake account detection tasks and compare their results with SHAP.

Acknowledgment

This work was supported by the Scientific Research and Innovation Support Fund in Jordan (project No. ICT-Ict/1/2021). We also want to thank Princess Sumaya University for Technology for their continued support.

References

[1] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[3] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[4] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, and S. Ioannidis, "Identification of twitter bots based on an explainable machine learning framework: the us 2020 elections case study," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 956–967.

[5] L. D. Samper-Escalante, O. Loyola-González, R. Monroy, and M. A. Medina-Pérez, "Bot datasets on twitter: Analysis and challenges," *Applied Sciences*, vol. 11, no. 9, p. 4105, 2021.

[6] W. Antoun, F. Baly, R. Achour, A. Hussein, and H. Hajj, "State of the art models for fake news detection tasks," in *2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT)*. IEEE, 2020, pp. 519–524.

[7] M. Mohammadrezaei, M. E. Shiri, A. M. Rahmani *et al.*, "Identifying fake accounts on social networks based on graph analysis and classification algorithms," *Security and Communication Networks*, vol. 2018, 2018.

[8] B. Erşahin, Ö. Aktaş, D. Kılınç, and C. Akyol, "Twitter fake account detection," in *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2017, pp. 388–392.

[9] S. Khaled, N. El-Tazi, and H. M. Mokhtar, "Detecting fake accounts on social media," in *2018 IEEE international conference on big data (big data)*. IEEE, 2018, pp. 3672–3681.

[10] N. Singh, T. Sharma, A. Thakral, and T. Choudhury, "Detection of fake profile in online social networks using machine learning," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2018, pp. 231–234.

[11] E. Van Der Walt and J. Eloff, "Using machine learning to detect fake identities: bots vs humans," *IEEE access*, vol. 6, pp. 6540–6549, 2018.

[12] M. B. Albayati and A. M. Altamimi, "Identifying fake facebook profiles using data mining techniques." *Journal of ICT Research & Applications*, vol. 13, no. 2, 2019.

[13] S. R. Sahoo and B. B. Gupta, "Fake profile detection in multimedia big data on online social networks," *International Journal of Information and Computer Security*, vol. 12, no. 2-3, pp. 303–331, 2020.

[14] R. Subhashini, R. Sethuraman, and B. K. Samhitha, "Prediction of fake instagram profiles using machine learning," *Annals of the Romanian Society for Cell Biology*, pp. 4490–4497, 2021.

[15] M. Alazab, S. KP, S. Srinivasan, S. Venkatraman, Q.-V. Pham *et al.*, "Deep learning for cyber security applications: A comprehensive survey," Tech. Rep., 2021.

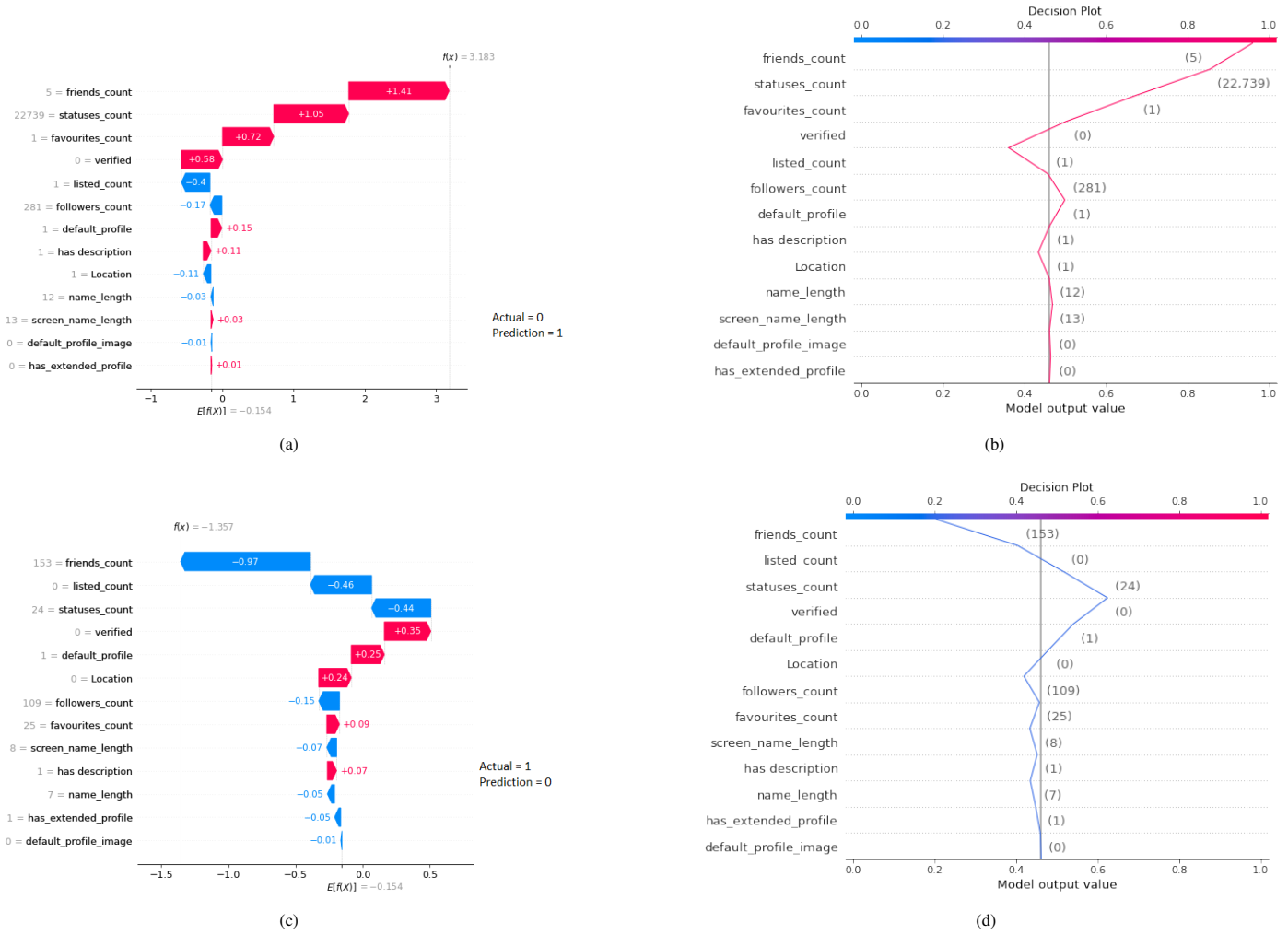

Fig. 10. Local bar plot / decision charts - Dataset2.

[16] B. Gulmezoglu, "Xai-based microarchitectural side-channel analysis for website fingerprinting attacks and defenses," *IEEE transactions on dependable and secure computing*, vol. 19, no. 6, pp. 4039–4051, 2021.

[17] M. Kouvela, I. Dimitriadis, and A. Vakali, "Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities," in *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, 2020, pp. 55–63.

[18] B. Bakhshandeh. Instagram fake spammer genuine accounts. [Online]. Available: https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts/data

[19] 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 *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 963–972.

[20] P. Kumbhar and M. Mali, "A survey on feature selection techniques and classification algorithms for efficient text classification," *International Journal of Science and Research*, vol. 5, no. 5, p. 9, 2016.

[21] N. Verma. (2022) Xgboost algorithm explained in less than 5 minutes. [Online]. Available: https://medium.com/@techynilesh/xgboost-algorithm-explained-in-less-than-5-minutes-b561dcc1ccee

[22] D. Rothman, *Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps*. Packt Publishing Ltd, 2020.

[23] R. LIN. Explainable ai with shap — income prediction example. [Online]. Available: https://reneelin2019.medium.com/explainable-ai-with-shap-income-prediction-example-3050c19a261b

[24] T. Carneiro, R. V. M. Da Nóbrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. Reboucas Filho, "Performance analysis of google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61 677–61 685, 2018.

[25] R. Younisse, A. Ahmad, and Q. Abu Al-Haija, "Explaining intrusion detection-based convolutional neural networks using shapley additive explanations (shap)," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 126, 2022.