

## RESEARCH ARTICLE

# Identification of Spambots and Fake Followers on Social Network via Interpretable AI-Based Machine Learning

DANISH JAVED<sup>1</sup>, NOOR ZAMAN JHANJHI<sup>1</sup>, (Member, IEEE), NAVID ALI KHAN<sup>1</sup>,  
SAYAN KUMAR RAY<sup>1</sup>, ARAFAT AL-DHAQM<sup>1</sup>, AND VICTOR R. KEBANDE<sup>2</sup>, (Member, IEEE)

<sup>1</sup>School of Computer Science (SCS), Taylor's University, Subang Jaya 47500, Malaysia

<sup>2</sup>Department of Computer Science, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

Corresponding authors: Noor Zaman Jhanjhi (noorzaman.jhanjhi@taylors.edu.my) and Victor R. Kebande (victor.kebande@bth.se)

This work was supported by the Blekinge Institute of Technology through Grant.

**ABSTRACT** Social networking platforms like X (Twitter) serve as hubs for open human interaction, but they are also increasingly infiltrated by automated accounts masquerading as human users. These bots often engage in activities such as spreading fake news and manipulating public opinion during politically sensitive times like elections. Most of the current bot detection methods rely on black-box algorithms, raising concerns about their transparency and practical usability. This study aims to address these limitations by developing a novel methodology for the detection of spambots and fake followers using annotated data. To this end, we propose an interpretable machine learning (ML) framework, leveraging multiple ML algorithms with hyperparameters optimized through cross-validation, to enhance the detection process. Furthermore, we analyze several features and provide a unique feature set that is optimized to offer excellent performance for bot detection. Moreover, we utilize multiple interpretable AI techniques which include Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP will help to display the effects of particular characteristics on the model's prediction which will help in determining whether an account is a bot or a legitimate user. LIME will help to comprehend the model's predictions, offering clarity regarding the traits or attributes that drive the classification conclusion. LIME allows researchers to detect bot-like activity in social networks by generating locally faithful explanations for each prediction. Our model offers enhanced interpretability by clearly highlighting the impact of various features used for spam and fake follower detection when compared to existing state-of-the-art social network bot detection systems. The results showcase the model's ability to identify key distinguishing attributes between bots and legitimate users which offers a transparent and effective solution for social network bot detection. Additionally, we utilize two comprehensive datasets including Cresci-15 and Cresci-17, which serve as robust baselines for comparison. Our model showcases its effectiveness by outperforming other methods while providing interpretability which increases performance and reliability for the task of bot detection.

**INDEX TERMS** Interpretable AI, social network, bot detection, fake followers, spambots.

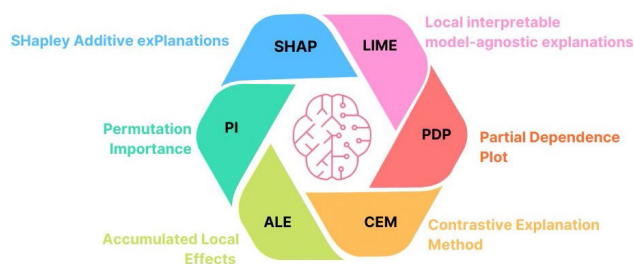
## I. INTRODUCTION

Social networks have become the key source of information in the new age of mankind. X formerly known as Twitter is presently among the most prevalent and widely used

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko<sup>1</sup>.

social media sites and thus it plays an important role in online conversations and helps connect millions of active users [1]. However, its substantial social and economic influence has also made it an attractive target for malicious actors seeking to manipulate and influence public opinion and decision-making. X has for some time been a prime target for automated programs, or "bots," due to its open

nature and expanding user base. These bots can be useful as legitimate bots produce a lot of educational tweets, such as blogs and news updates. Malicious bots, however, disseminate spam or harmful material. The characteristics used by current Twitter bot identification algorithms are often derived from user data, including timestamps, friendship, behavior, and network connection [2], [3]. Nevertheless, feature engineering requires a lot of work and effort. Social bots have the potential to facilitate the dissemination of misinformation, including fake news, rumors, and hate speech, by rapidly amplifying low-credibility content on X through interactions with high-profile users and strategic mentions [4]. Most of the aforementioned issues are controlled through the use of bots. A botnet is a collection of bots designed to execute specific tasks [5], while a Sybil account represents a fabricated identity that does not correspond to or originate from a real human user [6]. These botnets and Sybil accounts are frequently employed to amplify disinformation and disrupt genuine discourse, contributing to the challenges of maintaining integrity in online platforms.



**FIGURE 1.** Interpretable AI techniques.

Machine learning (ML) has been successfully utilized in a vast range of areas such as sports analytics [7], sentiment analysis [8], [9], fake news detection [10] and social bot detection [11]. Our study focuses on interpretable machine learning (XAI) as it has been used in different areas to improve performance and to gain better comprehension of the model. Figure 1 provides the most commonly used Interpretable AI techniques among which SHAP and LIME are the most popular. Interpretable ML provides insight into how a particular data point or data point affects the prediction model using a variety of methods such as factor analysis, local interpretation model-agnostic interpretation (LIME), and Shapley additive interpretation (SHAP) [12]. The added transparency helps users understand and trust AI systems while it also allows stakeholders to identify biases in these systems thus promoting accountability and fairness in AI applications. Overall, descriptive ML plays an important part in closing the disparity between AI algorithms and human comprehension which supports informed decision-making and increasing trust in AI technology. Thus, utilizing XAI for social network bot detection (SNBD) is an important step to gain a better understanding of its detection process [11].

Existing research utilizes various characteristics of the social network to differentiate between human and automated accounts. These features include user activity patterns (e.g., tweet frequency, timestamps), account metadata (e.g., follower/following ratios, account age), and social network structures (e.g., retweet and mention networks) [13], [14] etc. Supervised ML models and deep neural networks have been widely employed for this purpose [15], [16]. Traditional bot detection systems such as heuristic methods fail against evolving spambots, network-based approaches depend on narrow social networks, and earlier ML models employ limited characteristics, disregarding linguistic, temporal, and sentiment trends. Furthermore, the majority are not explainable, which makes it challenging to evaluate the data. Our Interpretable AI-based model addresses these gaps by integrating diverse feature sets. We enhance transparency with XAI which ensures improved accuracy, robustness, and interpretability. Furthermore, clustering and anomaly detection methods have been explored for unsupervised detection of anomalous behaviors linked to bots [17]. While these methods have shown promising results, they often lack scalability and adaptability due to their dependency on handcrafted feature engineering and static datasets. Moreover, the heavy reliance on black-box ML models limits their interpretability and creates barriers to understanding how decisions are made. Several challenges reduce the effectiveness of current bot detection methodologies. One of these challenges is feature engineering which is a labor-intensive process that requires domain expertise and manual effort to adapt the models to newer datasets and bots. Furthermore, bots exhibit dynamic and adaptive behavior through the evolution of their strategies to mimic human users more effectively and evade detection algorithms [11]. As a result, black-box detection models struggle to adapt to the constantly evolving nature of bot activities. Additionally, the lack of model interpretability in these methods undermines trust and transparency. Evaluation without interpretability is a challenge as we don't know if the model is identifying bots based on meaningful patterns or merely overfitting to noise in the data. Additionally, most methods are designed to optimize detection accuracy without considering the broader goals of generalizability and adaptability which are critical for real-world deployment on social networks. These gaps highlight the need for more transparent and interpretable detection frameworks. Therefore, the proposed methodology addresses these challenges by integrating interpretable machine learning (XAI) techniques into the bot detection process. These techniques boost the transparency of ML methods by providing insights into the contribution of individual characteristics to model predictions [18]. To that end, we offer the following contributions through our research. These contributions can help to progress the field of bot detection on social networks.

- This study presents an innovative interpretable bot-detection model built for detecting spambots and fake followers on social networking platforms, specifically Twitter/X. The model utilizes interpretable AI

techniques to offer clear and interpretable insights into the bot identification which improves the detection mechanism's credibility.

- This study delves into the numerous features of X and analyzes their influence on the bot detection model. Multiple explainable AI methods are utilized to study the behavior of different features within the context of bot detection to enhance the model's generalization capabilities through evaluation across a variety of bot types, through well-established datasets.
- This research validates the efficacy of XAI through enhanced performance for bot detection while providing greater transparency compared to other state-of-the-art methods. The proposed model attains superior detection results and offers insights into the model.

The remainder of the paper is structured as shown in Figure 2 below.

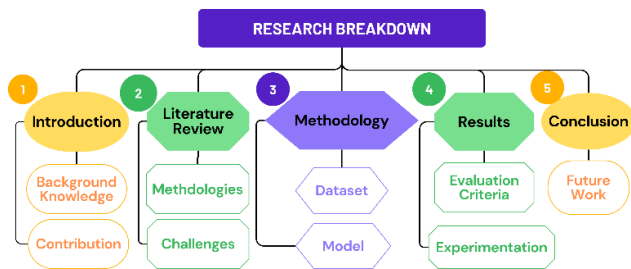


FIGURE 2. Research breakdown.

## II. LITERATURE REVIEW

The burgeoning interest in bot-detection challenges has precipitated a proliferation of academic inquiry, yielding a plethora of articles that proffer diverse methodologies. That notwithstanding, a gap exists in the extant literature, as the overwhelming majority of these approaches fail to provide transparent and interpretable results. In the subsequent sections, a concise review of prevailing bot detection strategies will be presented, accompanied by an examination of the challenges that necessitate further investigation. The preponderance of bot detection methodologies relies on supervised ML paradigms, which necessitate the utilization of one or multiple annotated datasets to train ML classifiers and develop an efficacious framework. These annotated datasets are frequently generated through human annotation, although alternative approaches such as leveraging pre-existing established models, crowdsourcing, or automated annotation techniques have also been employed to construct datasets for bot detection purposes. Table 1 highlights the key literature for interpretable AI-based bot detection. This literature discusses the purpose of each study and our findings on each research.

### A. SNBD METHODOLOGIES

In [23], the authors devised a novel approach by creating a corpus of honeypot accounts, specifically designed to attract spammer interactions, and subsequently logged

the corresponding profile information. This dataset was then augmented with a collection of regular user profiles, thereby enabling the development of a comprehensive classification algorithm that incorporates both user-centric and content-centric features. In another research [24], authors took a similar method, attempting to detect botnets that were run by the same person. Reference [25] employs crowdsourcing techniques for bot recognition on Facebook, and while it appeared to provide decent results; however, the inherent limitations of this method became apparent when the perpetual evolution and proliferation of bots rendered the approach increasingly unscalable, thereby underscoring the need for more adaptive and dynamic bot detection strategies. Crowdsourcing has been employed in several ways for data annotation jobs, which we shall discuss below.

The most common method [26], BotOrNot and subsequently Botometer, is based on the dataset supplied by [23], which has been augmented with new tweets for each identified account. The large number of distinct characteristics utilized to train the model was the approach's breakthrough. In [46], the authors provided a strategy for extracting this huge feature collection and confirmed their findings using a fresh annotated dataset. The findings verified the suggested model's efficiency while also highlighting unique shortcomings. For instance, the model's performance deteriorated when applied to the new dataset, as it was trained on earlier bot variants that exhibited distinct behavioral patterns and characteristics compared to the updated ones. The authors provide access to multiple labeled bot datasets, as referenced in [27], and demonstrate how the crowdfunding capabilities of the Botometer platform are leveraged to retrain the model and adapt to the evolving bot landscape. Alternatively, more straightforward yet efficacious methods for bot identification have been proposed named Stweeler [28], [29], which employs a click-bait strategy to gather user and tweet data for bot detection. Another technique [30] identifies automated accounts examining the unpredictability of the screen name, while another [31] demonstrates that the trained model is extremely efficient even with a thorough selection of 10 criteria. The majority of the papers that have been discussed employ simple ML algorithms, but other techniques employ Deep Learning (DL) or more complicated algorithms.

Reference [32] offers a DL-based method that utilizes neural networks to detect bots using a behavior-augmented model on users. Similarly, the authors of [23] suggest using an LSTM network that exploits the content and metadata of X combined with contextual user attributes to determine if a tweet was made by an automated account. Alternatively, [33] presents a completely different method of bot identification, emphasizing the need to recognize concerted attacks instead of lone individuals. The aforementioned strategies employ a diverse array of features and ML techniques, yet they seemingly fail to adequately address additional difficulties, which are described further in the next section. Figure 3 presents the most commonly used SNBD method.

**TABLE 1. Key literature for XAI-based bot detection.**

Cite	Technique	Purpose	Findings
[18]	Explainable AI, ML	The goal of this work is to propose the development of a method for identifying social media bots through labeled data. To do this, an XAI-based ML methodology is utilized, where the hyper-parameters are adjusted and results are validated via K-fold.	Authors employ SHAP to explicate ML model predictions by evaluating attribute significance with game theory Shapley scores.
[19]	Explainable deep graph neural network (GNN)	The authors present XG-BoT, which is a comprehensible deep GNN approach for detecting botnet nodes. The suggested approach comprises a botnet identifier and an explainer module. This approach is effective at detecting fraudulent botnet nodes in massive networks.	XG-BoT was tested using real-world botnet network graph data. It beats cutting-edge methods in terms of major evaluation parameters. Furthermore, the authors show that the explainer module can produce valuable explanations for automated network forensics.
[20]	Explainable machine learning	This work presents a unique, replicable, and reusable Twitter bot-identifying technique. The system employs an ML based methodology which involves hundreds of characteristics. The primary goal of the suggested method is to train and verify various cutting-edge ML models to achieve the best detection performance.	Authors utilize their own dataset collected from Twitter during the 2020 US Presidential Elections, and further investigation is performed on different Twitter datasets to show that the method is better in terms of bot identification accuracy.
[21]	Deep learning	A novel methodology is presented for identifying social bots on the Sina Weibo site that combines DL and active learning techniques. The method includes a complete set of 30 characteristics that are organized into four dimensions: metadata, interaction, content, and timing. In particular, this study adds nine novel traits, representing a considerable contribution to the discipline.	These added features enable the framework to distinguish between social bots and real users inside the Sina Weibo ecosystem, thus boosting the effectiveness of bot detection techniques.
[22]	Generative Adversarial Network (GAN)	Authors employ GAN to enrich the available data for training the cutting-edge textual bot detection approach. Despite its ability to enrich datasets with limited labeled samples, the original Sequence GAN has a known convergence issue.	The authors addressed the constraint of convergence by developing a revolutionary framework called GANBOT, which adapts the GAN principle. Authors connect the generator and classifier using an LSTM layer that serves as a common link among them.

## B. CHALLENGES OF SNBD

Despite the plethora of scientific endeavors that have yielded various methods for detecting online social bots, as indicated in the aforementioned studies, there are still many outstanding difficulties. Even though many SNBD approaches employ more than 1,000 attributes to train their method [26], it remains unclear whether increasing the number of features necessarily enhances model efficiency. Moreover, the authors of the [34] highlight the significant impact of utilizing an extensive feature set on the scalability of bot detection systems. Interestingly, they also note that employing various subsets of publicly available labeled datasets can enhance model generalizability, as observed in the same study. Notably, the performance of machine learning-based bot detection models varies across different datasets. Consequently, the accumulation of additional datasets is essential to ensure that our training data encompasses a comprehensive range of bot behavioral features. The same conclusion is

obtained from the [27] and [35], which goes to a finer-grained categorization of bots, giving distinct datasets for each sort of bot. As a result, one major difficulty in online social bot identification is determining what qualities genuinely constitute a social bot.

X bots are often used for malevolent objectives ranging from distributing fake news to propaganda and astroturfing [36], [37]. The writers of [38] examined 245,000 profiles on X between the 2016 US presidential election and the 2018 midterm elections, detecting around 31,000 bots. The authors of [39] conducted an exhaustive analysis of 43 million election-related tweets pertinent to the U.S. Congress investigation into Russian interference during the 2016 U.S. election campaigns. Their findings suggest that a significant proportion of users, specifically 4.9% of liberal and 6.2% of conservative were automated accounts. Notably, their approach achieved precision and recall scores exceeding 90%. The authors of [40] present an examination of German

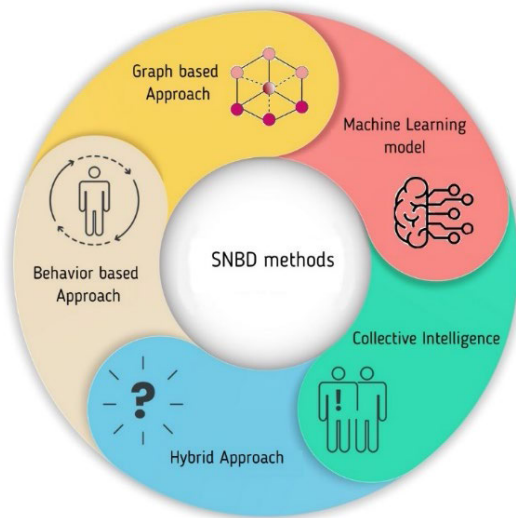


FIGURE 3. SNBD methods.

parties’ tweets from before and throughout the 2017 election cycle, demonstrating an increase in the use of social bots.

It is clear that Twitter bot identification is a difficult process that frequently needs thorough and robust treatment. Several ML-based methods, such as the BotOrNot [26] have been offered with a total of 1200 distinct characteristics combined with an ML classifier. An enhanced version of this system, Botometer, is detailed in [27], which needs X API keys to obtain user data during real-time calculations, making it inefficient to utilize real-time labeling tools in the case of large datasets. There is an increasing number of Twitter bot identification programs that use machine learning and data (statistical) analysis such as the Stweeler [28], the Debot [41], and the Retweet-Buster (RTbust) [42].

### III. METHODOLOGY

Our methodology employs interpretable AI-based machine learning to showcase a complete strategy for detecting spambots and phony followers on social media, assuring robustness, generalizability, and interpretability. We utilize a modular-based approach for the construction of our model where the process starts with data preprocessing of the dataset. We continue onwards to feature engineering where we extract several features and select the best attributes for bot detection. We utilize various types of attributes including user profile features, linguistic features, engagement features, and content-based features. In addition, we perform sentiment analysis on textual information such as tweet text and description text, and extract sentiment features. In the subsequent step, we partition the dataset into training, validation, and testing sets, ensuring that each train-test split maintains a consistent class ratio for both training and testing data through stratification. We perform extensive testing to compare the classification accuracy of bot versus regular users utilizing a variety of state-of-the-art ML algorithms and explainable AI approaches in order to create a reliable and accurate machine

learning-based bot recognition solution. We evaluate the performance of our model through a diverse range of ML-based algorithms and use K-fold cross-validation for results to avoid any bias in the model. Given that each machine learning method has a different set of parameters, it is crucial to use optimal parameters to determine which version of the classifier is best which allows for an equitable comparison. Figure 4 depicts our Interpretable AI-based model which utilizes a module-based methodology. Each module performs a specific task in order to optimize the task of spambots and fake follower identification. Further details on the methodology are presented in the subsections below.

#### A. DATASET

The Cresci-15 is a credible benchmark data for social network bot detection, introduced by [43]. It contains a combination of human and bot accounts, collected from Twitter, and is designed to evaluate the efficacy of bot detection algorithms. The dataset is composed of several subsets, each representing different bot types and human behaviors as shown in Table 2.

TABLE 2. Dataset characteristics (Cresci-15).

Sub-Dataset	Type	Accounts	Tweets
TFP (the fake project)	100% humans	469	563693
E13 (elections 2013)		1481	2068037
FSF (fastfollowerz)	100% fake followers	1169	22910
INT (intertwitter )		1337	58925
TWT (twittertechnology)		845	114192

The Cresci-17 dataset is a widely acknowledged standard dataset for bot detection on social media platforms, particularly Twitter [35]. This dataset is unique as it includes tweets from a variety of accounts, including genuine human users, typical social bots (made and run with the goal of deceiving people), and sophisticated social bots (more complex bots designed to mimic human behavior). The Cresci-17 dataset is used by the scientific community to build and validate bot detection methods, evaluate their accuracy and generalizability, and compare the performance of various methodologies. Its availability has made major contributions to improvements in the field of SNBD allowing for the creation of more effective tactics for recognizing and reducing bot influence on social networks. Table 3 provides the characteristics of the dataset.

#### B. DATA PREPROCESSING

Data preprocessing assumes paramount importance as it entails the transformation of raw data into a format that is useful for ML analysis [44]. This critical step enables the extraction of meaningful features which facilitates the development of accurate and reliable bot detection systems. We import libraries such as torch, torchtext, tqdm, emoji, and nltk to handle data preprocessing tasks. In our model, we utilize various features that were extracted from user

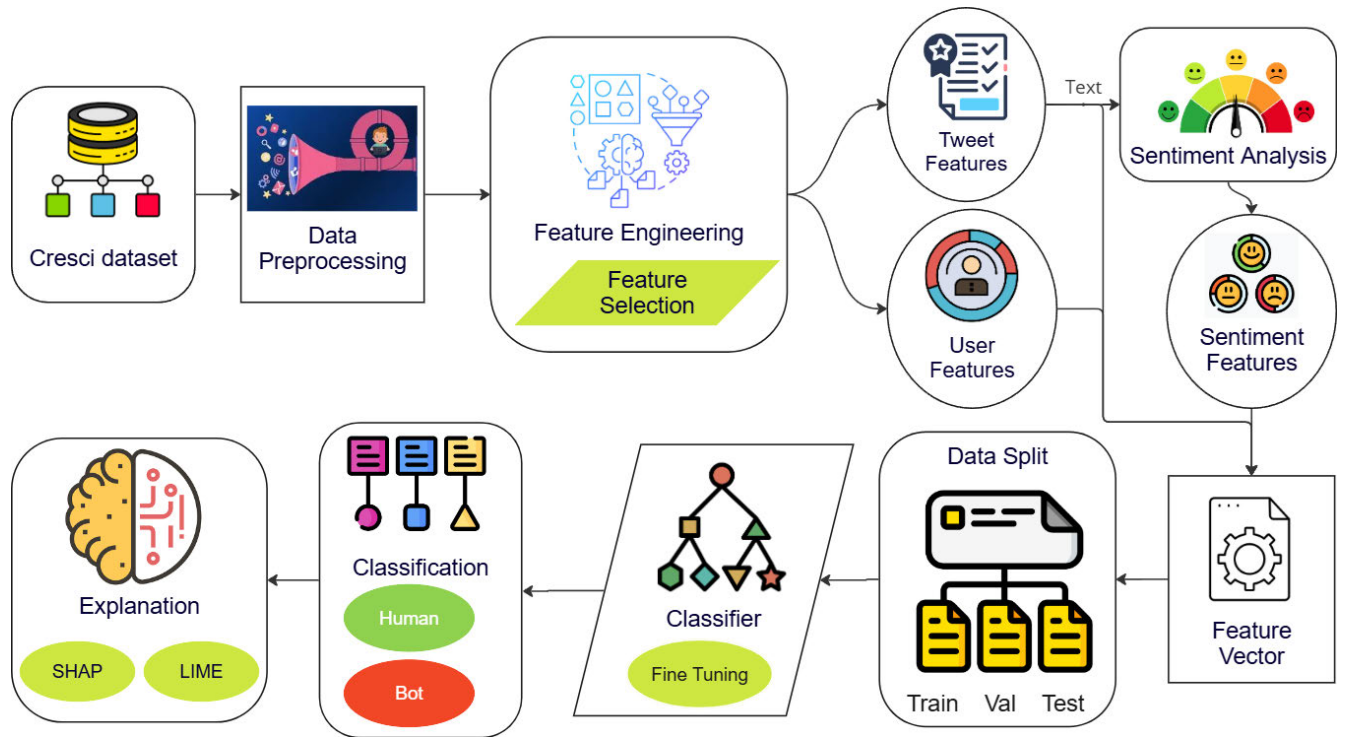


FIGURE 4. Proposed Model for identification of spambots and fake followers.

TABLE 3. Dataset characteristics (Cresci-17).

Bot Type	Description	Total accounts	Total tweets
Traditional spambots	Instances or examples of bots classified as spammers.	1000	145094
Social Spambots 1	Bots who retweet a political candidate from Italy.	991	1610176
Social Spambots 2	Bots who engage in spamming activities related to paid mobile applications.	3457	428542
Social Spambots 3	Bots who engage in spamming products are available for sale on Amazon.	464	1418626
Fake Followers	Fake profiles that follow the user.	3351	196027
Genuine accounts	Real human accounts that are authentic.	3474	8377522

profile information and tweet information. Figure 5 shows the top keywords within the description text of the real users while Figure 6 shows the top keywords for the bot users, offering insight into the content variety. As we can see it contains a lot of words that don't have any meaning, and some words are similar in both word clouds. Therefore, it is vital to preprocess the textual data in a way that allows the model to create a distinction between real user text and bot-generated text. The description feature is defined by textual information and the preparation for this feature includes a set of preprocessing steps that are performed within the feature

engineering pipeline. The description and tweet text data are essential for sentiment analysis as it allows us to extract sentiment-based features.

It is very important to deal with null values within the data because these null values can cause issues with tree-based classifiers such as Random Forest as these classifiers are equipped to handle null values. X account's description field may contain null values; therefore, the missing value imputations are substituted with a default value 'missing,' which indicates the lack of available data. However, the description\_length attribute for null descriptions is kept at zero. Raw Twitter data often contains noise, such as irrelevant symbols, URLs, mentions, and emojis. Preprocessing involves cleaning the text data by removing or replacing such elements. For example, emojis are converted into textual representations so that they can help with sentiment analysis. Special characters, punctuation, and whitespace are standardized or removed to ensure consistency and uniformity in the dataset, but this is only done for sentiment analysis as URL and punctuation information are used as features for our model. Stop words are commonly used but they are less informative and thus we remove them to minimize vocabulary size and prioritize words with significant informational substance.

### C. FEATURE SELECTION AND EXTRACTION

The dataset contains two distinct files which consist of user and tweet data. These two sets of data are utilized to extract different types of features as shown in Table 4. We investigated several features of X and developed numerous



column shows the features that contribute to the probability of an account being a ‘human’. The ‘Bot’ side of the column shows the positive indicators for the “Bot” classification. For example, a high number of replies correlates strongly with bot-like behavior due to automated engagement patterns to increase visibility. Bots frequently use hashtags to target specific audiences or trends which is shown by its contribution to the “bot” prediction with the value 0.11. Similarly, a low follower-following ratio is characteristic of bot accounts which indicates an imbalance in social reciprocity. On the other hand, retweet\_count of 0.61 is a medium level of retweet value which aligns with human-like sharing behavior. Figure 8 illustrates a 100% confidence prediction for “human”. We see the positive indicators for “Human” classification, for example, avg\_mentions and favorites\_count values are greater than zero. On the other hand, features such as default\_profile and low avg\_hashtag use marginally align with bot-like behavior.

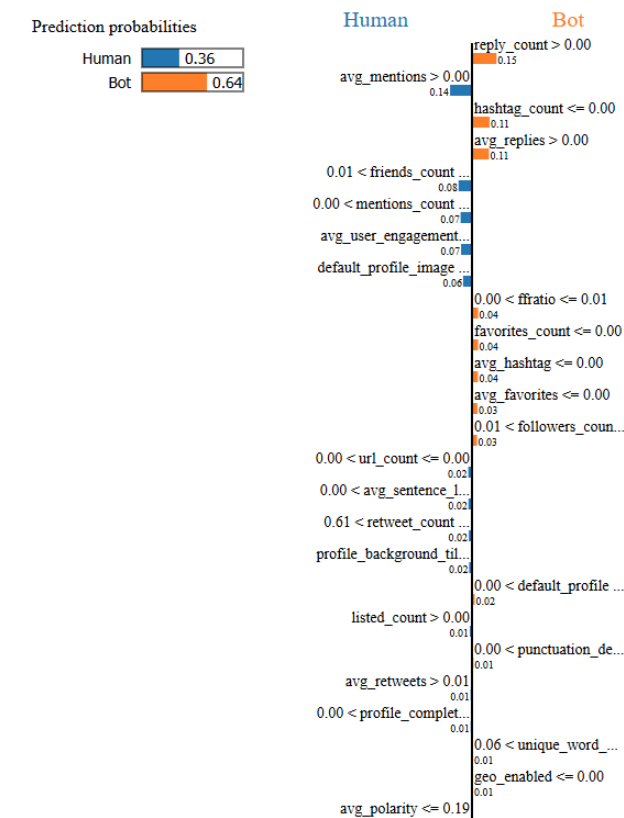


FIGURE 7. Lime explanation for “bot” prediction (cresci-15).

**E. EXPLAINABLE METHOD: SHAP**

SHAP is a technique that is used to enhance the interpretability of ML models in Twitter bot detection by providing insights into the contribution of individual features to the model’s predictions. SHAP is rooted in game theory and computes Shapley values which quantify the marginal contribution of each feature to the prediction by considering all possible combinations of feature subsets. This method

allows for a transparent and model-agnostic explanation of how features influence the decision-making process of the classifier. In our research, SHAP was employed to analyze and rank the most critical features influencing the predictions. Figures 9 show the visualization for the cresci-15 dataset while Figure 10 shows the SHAP values for the cresci-17 dataset.

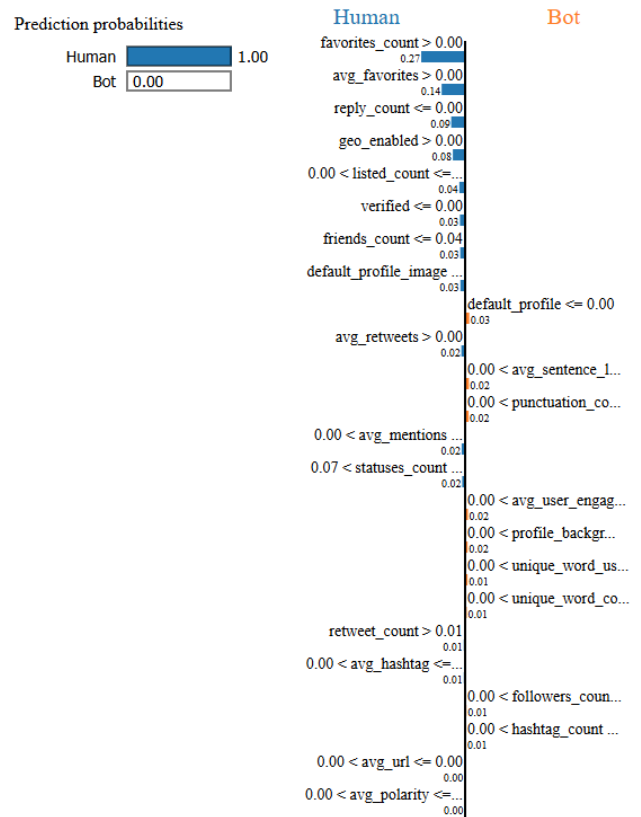


FIGURE 8. Lime explanation for “human” prediction (cresci-17).

The top twenty characteristics with the greatest influence on the output of the ML model are displayed. One point refers to a single Twitter user for each feature. The position of a point on the x-axis is the actual SHAP values which show the influence of an attribute on the model’s output for that particular X user. Mathematically, this corresponds to the probability or likelihood of a user engaging in harmful behavior compared to other Twitter users. In this context, a higher SHAP value indicates that a Twitter user is more likely to exhibit malicious behavior relative to a user with a lower SHAP value. The relevance of features is determined by the mean of their absolute Shapley values which is shown along the y-axis. Essentially, the SHAP values quantify the contribution of specific features such as ‘favorites\_count’ and ‘avg\_sentence\_length are high-value features that contribute toward the prediction as shown in Figure 9. This probabilistic interpretation helps in ranking users based on their potential for malicious behavior which allows for more targeted interventions in detecting and mitigating harmful actions on the platform.

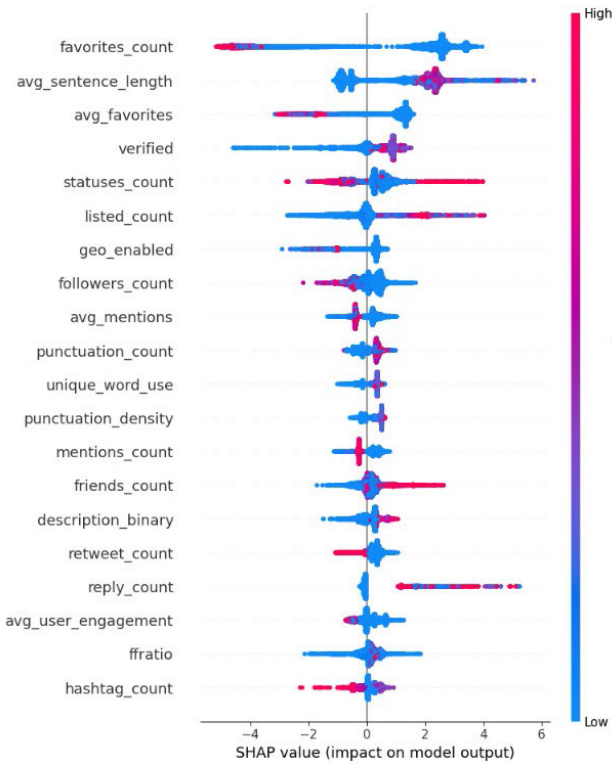


FIGURE 9. SHAP value for the cresci-15 dataset.

IV. RESULTS

In this section, we evaluate the performance of the proposed SNBD approach concerning its accuracy and generalizability in detecting bots and distinguishing them from normal users. The analysis focuses on assessing how effectively the model classifies users into bot or normal categories and provides a comprehensive understanding of its ability to detect bot accounts. We start the process with ‘shuffling’ which is a preprocessing step in ML where the order of the dataset is randomized. This process ensures that the data is not biased by any inherent order such as chronological arrangement or class grouping, which could otherwise affect the training and testing phases. Furthermore, we divide the data into 75%-25% for training and testing respectively.

A. EVALUATION CRITERIA

We evaluate the effectiveness of our interpretable ML-based spambot and fake follower detection model using a few key metrics:

*Accuracy:* It measures how accurate it is to correctly classify an account as a spam bot or legitimate user i.e., the overall effectiveness of the system’s detection.

*F1 measure:* The F-measure is a widely adopted metric in classification tasks, utilized to assess the overall performance of a model. It is calculated as the harmonic mean of precision and recall, combining these two important measures into a single value, as illustrated in Equation 1. Precision measures the accuracy of the model’s positive predictions, indicating the proportion of correctly identified positive instances (e.g.,

correctly classified bots) among all instances predicted as positive. It is calculated as the ratio of true positives to the sum of true positives and negatives.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

*Interpretability:* It is used to evaluate the interpretation of predictive models using techniques such as SHAP and LIME. This metric evaluates the model’s ability to provide an understandable explanation for its decisions, allowing participants to understand what is driving the distribution.

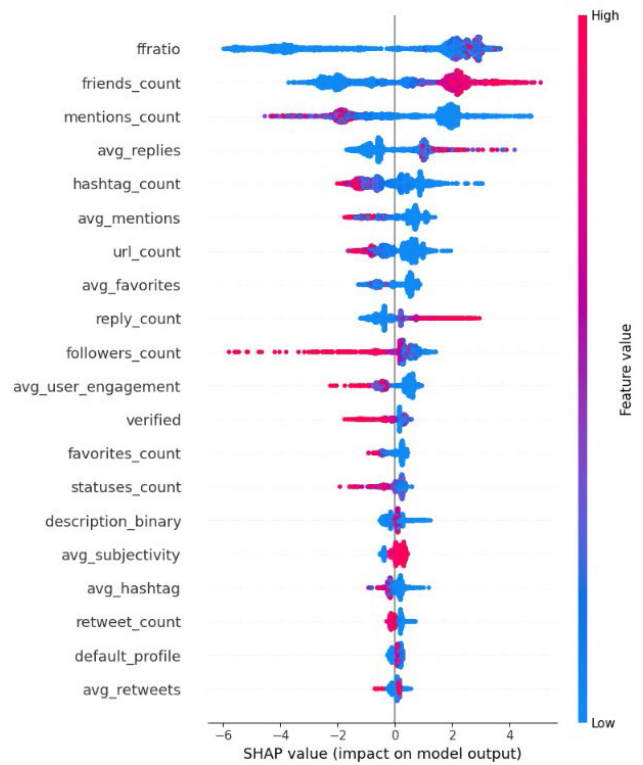


FIGURE 10. SHAP explanation for cresci-17 dataset.

*Area Under the Curve:* It measures the trade-off between the true positive rate (sensitivity) and the false positive rate. Its value ranges from 0 to 1, where 1 indicates perfect classification, 0.5 represents random guessing, and values closer to 0 suggest poor performance.

*K-Fold cross-validation:* In the context of identifying spam bots and fake followers on social media using explainable AI-based ML. The data was evaluated using various ML classifiers using 5-fold cross-validation and the 70%-30% retention method respectively. In 5-fold cross-validation, where the data set is divided into five groups, one group is used as the test set, and the others are used as the training set.

B. CLASSIFICATION RESULTS

In this section, we discuss the classification results obtained on two datasets mentioned in section III-A. We tested various machine-learning classifiers on these datasets to create

**TABLE 5. Results on the cresci-15 dataset.**

Classifiers	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	0.990	0.994	0.990	0.992	0.999
SVM	0.959	0.972	0.962	0.967	0.983
Decision Tree	0.976	0.980	0.982	0.981	0.974
XGBoost	0.991	0.994	0.991	0.993	0.999
LightGBM	<u>0.991</u>	<u>0.994</u>	<u>0.992</u>	<u>0.993</u>	<u>0.999</u>
Logistic Regression	0.954	0.973	0.953	0.963	0.977
Extra Trees	0.987	0.994	0.986	0.990	0.999
Naïve Bayes	0.768	0.739	0.980	0.842	0.966
AdaBoost	0.986	0.991	0.988	0.990	0.998

a more authentic model for bot detection. The results presented in Table 5 and Table 6 are obtained through K-fold cross-validation with the value set to 5. This removes the issue of overfitting and provides repeatable results. We show the results with respect to accuracy, precision, recall, F1 score, and AUC for a more comprehensive understanding of the results. Table 5 shows the results for the cresci-15 dataset where multiple classifiers were tested among which LightGBM provides the best accuracy and F1 of 0.991 and 0.993 respectively, but falls slightly behind in recall. Table 6 shows the results for the cresci-17 dataset where multiple classifiers were tested among which XGBoost and LightGBM provide the best accuracy and an F1 score of 0.990 and 0.993. All classifiers perform competitively except Naive Bayes and SVM which provide significant reduction in accuracy and F1. It is important to analyze the trade-offs between precision and recall for each classifier to better understand their ability to handle false positives. Minimizing false positives is a crucial task since we want to avoid unnecessarily blocking the accounts of genuine users. Our model minimizes false positives which is evident from the results where our model has provided high precision across various types of bots and datasets.

## V. DISCUSSION AND COMPARISON

Our results show the effectiveness of interpretable ML-based models for the identification of spam bots and fake followers on social network platforms such as X. Our study focused on addressing the limitations of current bot detection methods by developing an interpretable model where we utilize techniques such as SHAP and LIME. The use of these techniques enhances our understanding of the decision model by providing insight into the significance and interpretation of the features used in SNBD. For example, SHAP analysis shows that factors such as favorites\_count, avg\_mentions, unique\_word\_use, and followers-following ratio are high-value features that influence the prediction model. This transparency is important for building trust in these methods as it allows models to be accurate and precise. This research addresses several important issues in the task of SNBD. First, it reduces the black-box nature of traditional bot detection

methods and provides clear reasons and explanations behind the contribution of features using XAI. The use of XAI can cause computational overhead which presents a significant challenge as the datasets such as Cresci-15 and Cresci-17 are characterized by high dimensionality and large sample sizes. SHAP's reliance on approximating shapley values introduces exponential time complexity which can become computationally expensive when dealing with huge feature sets. We deal with this issue by presenting a compact feature set that utilizes thirty-one features to provide competitive results. Similarly, LIME's requirement to train local surrogate models for each prediction increases runtime in scenarios that involve vast amounts of data. But LIME may not be as costly if we are just looking for a limited number of instance prediction results. These computational demands may limit the deployment of XAI in large-scale frameworks. These issues can be addressed through the utilization of optimization strategies such as dimensionality reduction thus reducing the total number of features while still maintaining high performance. Ensuring the generalizability of our SNBD model is critical and thus we train our model on two large datasets that include various sorts of bot accounts such as traditional spambots, social spambots, and fake followers. Our model learns to detect and discriminate between distinct types of bots and genuine users with high accuracy and reliability.

The non-interpretability of models in bot detection on social networks imposes several limitations. The lack of transparency in black-box models undermines trust because they provide predictions without explaining the rationale behind their decisions. This makes it challenging to justify classification outcomes. Furthermore, the absence of interpretability complicates debugging which makes it difficult to identify and correct misclassifications such as false positives or negatives. Additionally, non-interpretability models are prone to biases which are present in the training data which leads to discriminatory outcomes that are difficult to detect and mitigate without insight into feature contributions. Moreover, bots on social networks often exhibit dynamic and evolving behavior to evade detection, and non-interpretability hinders the ability of models to adapt effectively to such changes. This rigidity can result in a reliance on static features that may lose relevance over time. Finally,

**TABLE 6. Results for cresci-17 dataset.**

Classifiers	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	0.988	0.992	0.992	0.992	0.998
SVM	0.964	0.981	0.972	0.977	0.991
Decision Tree	0.984	0.989	0.989	0.989	0.978
XGBoost	<u>0.990</u>	<u>0.994</u>	<u>0.993</u>	<u>0.993</u>	<u>0.999</u>
LightGBM	<u>0.990</u>	<u>0.994</u>	<u>0.993</u>	<u>0.993</u>	<u>0.999</u>
Logistic Regression	0.939	0.964	0.955	0.981	0.981
Extra Trees	0.987	0.991	0.992	0.991	0.998
Naïve Bayes	0.919	0.992	0.899	0.944	0.979
AdaBoost	0.983	0.989	0.990	0.989	0.997

**TABLE 7. Result comparison with baselines (Cresci-15).**

Cite	Accuracy	F1
[50]	0.985	0.988
[51]	0.978	0.980
[52]	0.988	0.988
[53]	0.977	0.975
[54]	0.972	0.978
[55]	0.983	0.987
Ours (LightGBM)	<u>0.991</u>	<u>0.993</u>

**TABLE 8. Result comparison with baselines (Cresci-17).**

Cite	Accuracy	F1
[15]	0.980	0.964
[34]	0.985	0.989
[56]	0.982	0.977
[57]	0.956	0.967
[58]	0.967	0.977
Ours (XGBoost)	<u>0.990</u>	<u>0.993</u>

non-interpretability limits the ability to gain insights into feature relevance which restricts opportunities for refining models and identifying key behavioral indications of bots. Addressing these limitations through XAI techniques significantly enhances model transparency and adaptability which allows for more effective bot detection on platforms like X.

Table 7 and Table 8 show the comparison between different studies on the cresci-15 and cresci-17 dataset. The majority of the baseline results were collected from [49] and [50] while others were collected through manual inspection. The results demonstrate the superior performance of our proposed models for social network bot detection. For the cresci-15 dataset, LightGBM performed the best showcasing an accuracy of 0.991 and F1 score of 0.993. For the Cresci-17 dataset, XGBoost achieved an accuracy of 0.990, and an F1-score of 0.993, the XGBoost model outperforms all existing state-of-the-art methods in precision, recall, and

F1. This highlights its ability to minimize false positives and accurately identify bot accounts, achieving a strong balance between precision and recall. As shown in Table 4, our method uses a wide range of rich features that capture different facets of sentiment analysis, language patterns, content attributes, and user behavior. In contrast to conventional bot detection models that depend on a small number of network-based or profile-based characteristics, our feature engineering offers a multifaceted strategy that greatly enhances classification performance. Furthermore, the incorporation of XAI in bot detection helps to select the best features that provide better performance compared to previous models.

## VI. CONCLUSION

This research presents a unique way to differentiate between bots and real users on X by using an interpretable ML framework that extracts and analyzes attributes for the task of SNBD. The proposed methodology involves the extraction of a diverse set of features derived from the datasets discussed in Section III-A. The model was trained on various features that were finalized through explainable AI techniques to improve the detection of social and spam bots as well as fake followers. This approach increased the accuracy and reliability of our model and gave important insights into potential patterns which enhanced transparency for social security. This is done through the incorporation of the XAI techniques SHAP and LIME into the model which allows the researchers to understand the impact of the features on the model. This information allowed us to reduce the size of the feature set to include the most important features which reduced the workload for the ML model. The significance of this study lies in its ability to bridge the gap between model accuracy and transparency thus addressing the key challenges in bot detection by offering an interpretable methodology. This approach improves the reliability of detection models and provides actionable insights into feature relevance which paves the way for more efficient bot detection. Our model still has limitations due to its reliance on utilizing a specific feature set. This can be challenging when dealing with new bots. Detecting new-generation bots that mimic human activity remains a critical challenge because the bots continue

to evolve with more sophisticated behaviors. Future research could investigate adaptive models capable of learning from evolving bot behaviors, incorporating continuous learning mechanisms to stay ahead of these advancements. Given the interaction-based nature of social networks, graph neural networks could be integrated to enhance feature representation and extraction. Future research should explore combining graph-based representations with explainable AI to provide deeper insights into network behaviors and bot detection.

## CONTRIBUTIONS

Conceptualization, Danish Javed; Formal analysis, Danish Javed, Noor Zaman Jhanjhi; Funding acquisition, Sayan Kumar Ray, Arafat Al-Dhaqm, Victor R. Kebande; Investigation, Danish Javed; Methodology, Danish Javed; Project administration, Noor Zaman Jhanjhi and Navid Ali Khan; Resources, Sayan Kumar Ray, Arafat Al-Dhaqm, Victor R. Kebande; Validation, Danish Javed, Noor Zaman Jhanjhi; Writing – original draft, Danish Javed; Writing – review & editing, Danish Javed, Sayan Kumar Ray, Noor Zaman Jhanjhi, Navid Ali Khan.

## REFERENCES

- [1] E. Cano-Marin, M. Mora-Cantalops, and S. Sánchez-Alonso, "Twitter as a predictive system: A systematic literature review," *J. Bus. Res.*, vol. 157, Mar. 2023, Art. no. 113561, doi: [10.1016/j.jbusres.2022.113561](https://doi.org/10.1016/j.jbusres.2022.113561).
- [2] F. Tabassum, S. Mubarak, L. Liu, and J. T. Du, "How many features do we need to identify Bots on Twitter?" in *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, and R. D. Frank, Eds., Cham, Switzerland: Springer, 2023, pp. 312–327.
- [3] R. Al-Azawi and S. O. AL-Mamory, "Feature extractions and selection of bot detection on Twitter a systematic literature review," *Inteligencia Artif.*, vol. 25, no. 69, pp. 57–86, Apr. 2022, doi: [10.4114/intartif.vol25iss69pp57-86](https://doi.org/10.4114/intartif.vol25iss69pp57-86).
- [4] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025, doi: [10.1016/j.ipm.2019.03.004](https://doi.org/10.1016/j.ipm.2019.03.004).
- [5] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Design and analysis of a social botnet," *Comput. Netw.*, vol. 57, no. 2, pp. 556–578, Feb. 2013, doi: [10.1016/j.comnet.2012.06.006](https://doi.org/10.1016/j.comnet.2012.06.006).
- [6] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network Sybils in the wild," *ACM Trans. Knowl. Discovery from Data*, vol. 8, no. 1, pp. 1–29, Feb. 2014, doi: [10.1145/2556609](https://doi.org/10.1145/2556609).
- [7] D. Javed, N. Z. Jhanjhi, and N. A. Khan, "Football analytics for goal prediction to assess player performance," in *Proc. Int. Conf. Innov. Technol. Sports (RevealDNA ICITS)*, Apr. 2023, pp. 245–257, doi: [10.1007/978-981-99-0297-2\\_20](https://doi.org/10.1007/978-981-99-0297-2_20).
- [8] M. Humayun, D. Javed, N. Jhanjhi, M. F. Almufareh, and S. N. Almuayqil, "Deep learning based sentiment analysis of COVID-19 tweets via resampling and label analysis," *Comput. Syst. Sci. Eng.*, vol. 47, no. 1, pp. 575–591, 2023.
- [9] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for improved sentiment analysis via random minority oversampling for user tweet review classification," *Electronics*, vol. 11, no. 19, p. 3058, Sep. 2022, doi: [10.3390/electronics11193058](https://doi.org/10.3390/electronics11193058).
- [10] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi, "A hybrid transformer-based model for optimizing fake news detection," *IEEE Access*, vol. 12, pp. 160822–160834, 2024, doi: [10.1109/ACCESS.2024.3476432](https://doi.org/10.1109/ACCESS.2024.3476432).
- [11] D. Javed, N. Jhanjhi, N. A. Khan, S. K. Ray, A. A. Mazroa, F. Ashfaq, and S. R. Das, "Towards the future of bot detection: A comprehensive taxonomical review and challenges on Twitter/X," *Comput. Netw.*, vol. 254, Dec. 2024, Art. no. 110808, doi: [10.1016/j.comnet.2024.110808](https://doi.org/10.1016/j.comnet.2024.110808).
- [12] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates Inc., Jan. 2017, pp. 4768–4777.
- [13] M. Aljabri, R. Zagrouba, A. Shahid, F. Alnasser, A. Saleh, and D. M. Alomari, "Machine learning-based social media bot detection: A comprehensive literature review," *Social Netw. Anal. Mining*, vol. 13, no. 1, pp. 1–40, Jan. 2023, doi: [10.1007/s13278-022-01020-5](https://doi.org/10.1007/s13278-022-01020-5).
- [14] K. Hayawi, S. Saha, M. M. Masud, S. S. Mathew, and M. Kaosar, "Social media bot detection with deep learning methods: A systematic review," *Neural Comput. Appl.*, vol. 35, no. 12, pp. 8903–8918, Mar. 2023, doi: [10.1007/s00521-023-08352-z](https://doi.org/10.1007/s00521-023-08352-z).
- [15] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci.*, vol. 467, pp. 312–322, Oct. 2018, doi: [10.1016/j.ins.2018.08.019](https://doi.org/10.1016/j.ins.2018.08.019).
- [16] A. Dehghan, K. Siuta, A. Skorupka, A. Dubey, A. Betlen, D. Miller, W. Xu, B. Kamiński, and P. Prałat, "Detecting bots in social-networks using node and structural embeddings," *J. Big Data*, vol. 10, no. 1, pp. 1–37, Jul. 2023, doi: [10.1186/s40537-023-00796-3](https://doi.org/10.1186/s40537-023-00796-3).
- [17] H. Peng, J. Zhang, X. Huang, Z. Hao, A. Li, Z. Yu, and P. S. Yu, "Unsupervised social bot detection via structural information theory," *ACM Trans. Inf. Syst.*, vol. 42, no. 6, pp. 1–42, Nov. 2024, doi: [10.1145/3660522](https://doi.org/10.1145/3660522).
- [18] D. Javed, N. Z. Jhanjhi, and N. A. Khan, "Explainable Twitter bot detection model for limited features," in *Proc. Int. Conf. Green Energy, Comput. Intell. Technol. (GEN-CITY)*, Aug. 2023, Sep. 2023, pp. 476–481, doi: [10.1049/icp.2023.1822](https://doi.org/10.1049/icp.2023.1822).
- [19] W. W. Lo, G. Kulatilleke, M. Sarhan, S. Layeghy, and M. Portmann, "XG-BoT: An explainable deep graph neural network for botnet detection and forensics," *Internet Things*, vol. 22, Jul. 2023, Art. no. 100747, doi: [10.1016/j.iot.2023.100747](https://doi.org/10.1016/j.iot.2023.100747).
- [20] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, and S. Ioannidis, "Explainable machine learning pipeline for Twitter bot detection during the 2020 U.S. presidential elections," *Softw. Impacts*, vol. 13, Aug. 2022, Art. no. 100333, doi: [10.1016/j.simpa.2022.100333](https://doi.org/10.1016/j.simpa.2022.100333).
- [21] Y. Wu, Y. Fang, S. Shang, J. Jin, L. Wei, and H. Wang, "A novel framework for detecting social bots with deep neural networks and active learning," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106525, doi: [10.1016/j.knsys.2020.106525](https://doi.org/10.1016/j.knsys.2020.106525).
- [22] S. Najari, M. Salehi, and R. Farahbakhsh, "GANBOT: A GAN-based framework for social bot detection," *Social Netw. Anal. Mining*, vol. 12, no. 1, Dec. 2022, Art. no. 4, doi: [10.1007/s13278-021-00800-9](https://doi.org/10.1007/s13278-021-00800-9).
- [23] K. Lee, B. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, Aug. 2021, vol. 5, no. 1, pp. 185–192, doi: [10.1609/icwsm.v5i1.14106](https://doi.org/10.1609/icwsm.v5i1.14106).
- [24] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proc. 26th Annu. Comput. Secur. Appl. Conf.* New York, NY, USA: Association for Computing Machinery, Dec. 2010, pp. 1–9, doi: [10.1145/1920261.1920263](https://doi.org/10.1145/1920261.1920263).
- [25] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao, "Social Turing tests: Crowdsourcing Sybil detection," 2012, *arXiv:1205.3856*.
- [26] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "BotOrNot: A system to evaluate social bots," in *Proc. 25th Int. Conf. Companion World Wide Web WWW Companion*, Apr. 2016, pp. 273–274, doi: [10.1145/2872518.2889302](https://doi.org/10.1145/2872518.2889302).
- [27] K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," *Human Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48–61, Jan. 2019, doi: [10.1002/hbe2.115](https://doi.org/10.1002/hbe2.115).
- [28] Z. Gilani, L. Wang, J. Crowcroft, M. Almeida, and R. Farahbakhsh, "Stweeler: A framework for Twitter bot analysis," in *Proc. 25th Int. Conf. Companion World Wide Web WWW Companion*, Jul. 2016, pp. 37–38, doi: [10.1145/2872518.2889360](https://doi.org/10.1145/2872518.2889360).
- [29] Z. Gilani, R. Farahbakhsh, and J. Crowcroft, "Do bots impact Twitter activity?" in *Proc. 26th Int. Conf. World Wide Web Companion*, Apr. 2017, pp. 781–782, doi: [10.1145/3041021.3054255](https://doi.org/10.1145/3041021.3054255).
- [30] D. M. Beskow and K. M. Carley, "Its all in a name: Detecting and labeling bots by their name," *Comput. Math. Org. Theory*, vol. 25, no. 1, pp. 24–35, Mar. 2019, doi: [10.1007/s10588-018-09290-1](https://doi.org/10.1007/s10588-018-09290-1).
- [31] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 French presidential election," *1st Monday*, vol. 22, no. 8, Jul. 2017, doi: [10.5210/fm.v22i8.8005](https://doi.org/10.5210/fm.v22i8.8005).
- [32] C. Cai, L. Li, and D. Zengi, "Behavior enhanced deep bot detection in social media," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 128–130, doi: [10.1109/ISI.2017.8004887](https://doi.org/10.1109/ISI.2017.8004887).
- [33] C. Grimme, D. Assenmacher, and L. Adam, "Changing perspectives: Is it sufficient to detect social bots?" in *Social Computing and Social Media. User Experience and Behavior*, G. Meiselwitz, Ed., Cham, Switzerland: Springer, 2018, pp. 445–461.

- [34] K. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and generalizable social bot detection through data selection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 1, pp. 1096–1103.
- [35] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proc. 26th Int. Conf. World Wide Web Companion WWW Companion*, Jan. 2017, pp. 963–972, doi: [10.1145/3041021.3055135](https://doi.org/10.1145/3041021.3055135).
- [36] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," 2019, *arXiv:1901.06437*.
- [37] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 U.S. presidential election," *Nature Commun.*, vol. 10, no. 1, Jan. 2019, Art. no. 7, doi: [10.1038/s41467-018-07761-2](https://doi.org/10.1038/s41467-018-07761-2).
- [38] L. Luceri, A. Deb, S. Giordano, and E. Ferrara, "Evolution of bot and human behavior during elections," *1st Monday*, vol. 24, Aug. 2019, Art. no. 9, doi: [10.5210/fm.v24i9.10213](https://doi.org/10.5210/fm.v24i9.10213).
- [39] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 258–265, doi: [10.1109/ASONAM.2018.8508646](https://doi.org/10.1109/ASONAM.2018.8508646).
- [40] T. R. Keller and U. Klinger, "Social bots in election campaigns: Theoretical, empirical, and methodological implications," *Political Commun.*, vol. 36, no. 1, pp. 171–189, Jan. 2019, doi: [10.1080/10584609.2018.1526238](https://doi.org/10.1080/10584609.2018.1526238).
- [41] N. Chavoshi, H. Hamooni, and A. Mueen, "DeBot: Twitter bot detection via warped correlation," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 817–822, doi: [10.1109/ICDM.2016.0096](https://doi.org/10.1109/ICDM.2016.0096).
- [42] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "RTbust: Exploiting temporal patterns for botnet detection on Twitter," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 183–192.
- [43] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake Twitter followers," *Decis. Support Syst.*, vol. 80, pp. 56–71, Dec. 2015, doi: [10.1016/j.dss.2015.09.003](https://doi.org/10.1016/j.dss.2015.09.003).
- [44] M. F. Almuftareh, N. Jhanjhi, N. A. Khan, S. N. Almuayqil, M. Humayun, and D. Javed, "BERTSent: Transformer-based model for sentiment analysis of penta-class tweet classification," *IEEE Access*, vol. 12, pp. 196803–196817, 2024, doi: [10.1109/ACCESS.2024.3515836](https://doi.org/10.1109/ACCESS.2024.3515836).
- [45] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted Twitter spam," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 4, pp. 914–925, Apr. 2017, doi: [10.1109/TIFS.2016.2621888](https://doi.org/10.1109/TIFS.2016.2621888).
- [46] J. V. F. Abreu, C. G. Ralha, and J. J. C. Gondim, "Twitter bot detection with reduced feature set," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2020, pp. 1–6, doi: [10.1109/ISI5.525](https://doi.org/10.1109/ISI5.525).
- [47] J. Shin, "Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: Comparative fNIRS study," *Biomed. Eng. Lett.*, vol. 13, no. 4, pp. 689–703, Nov. 2023, doi: [10.1007/s13534-023-00291-x](https://doi.org/10.1007/s13534-023-00291-x).
- [48] R. S. Tiwari, "Hate speech detection using LSTM and explanation by LIME (local interpretable model-agnostic explanations)," in *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*. San Mateo, CA, USA: Morgan Kaufmann, 2024, pp. 93–110, doi: [10.1016/B978-0-443-22009-8.00005-7](https://doi.org/10.1016/B978-0-443-22009-8.00005-7).
- [49] S. Feng, H. Wan, N. Wang, J. Li, and M. Luo, "TwiBot-20: A comprehensive Twitter bot detection benchmark," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.*, Jun. 2021, pp. 4485–4494, doi: [10.1145/3459637.3482019](https://doi.org/10.1145/3459637.3482019).
- [50] Y. Liu, Z. Tan, H. Wang, S. Feng, Q. Zheng, and M. Luo, "BotMoE: Twitter bot detection with community-aware mixtures of modal-specific experts," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2023, pp. 485–495, doi: [10.1145/3539618.3591646](https://doi.org/10.1145/3539618.3591646).
- [51] Z. Yu, L. Bai, O. Ye, and X. Cong, "Social robot detection method with improved graph neural networks," *Comput. Mater. Continua*, vol. 78, no. 2, pp. 1773–1795, 2024, doi: [10.32604/cmc.2023.047130](https://doi.org/10.32604/cmc.2023.047130).
- [52] S. Li, B. Qiao, K. Li, Q. Lu, M. Lin, and W. Zhou, "Multi-modal social bot detection: Learning homophilic and heterophilic connections adaptively," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3908–3916, doi: [10.1145/3581783.3612569](https://doi.org/10.1145/3581783.3612569).
- [53] S. Shi, J. Chen, Z. Wang, Y. Zhang, Y. Zhang, C.-Q. Fu, K. Qiao, and B. Yan, "SStackGNN: Graph data augmentation simplified stacking graph neural network for Twitter bot detection," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, pp. 1–13, Apr. 2024, doi: [10.1007/s44196-024-00496-7](https://doi.org/10.1007/s44196-024-00496-7).
- [54] S. Feng, Z. Tan, R. Li, and M. Luo, "Heterogeneity-aware Twitter bot detection with relational graph transformers," in *Proc. 36th AAAI Conf. Artif. Intell. (AAAI-22)*, Jan. 2021, pp. 3977–3985. Accessed: Mar. 12, 2023.
- [55] Z. Lei, H. Wan, W. Zhang, S. Feng, Z. Chen, J. Li, Q. Zheng, and M. Luo, "BIC: Twitter bot detection with text-graph interaction and semantic consistency," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2023, pp. 10326–10340.
- [56] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "DNA-inspired online behavioral modeling and its application to spambot detection," *IEEE Intell. Syst.*, vol. 31, no. 5, pp. 58–64, Sep. 2016, doi: [10.1109/MIS.2016.29](https://doi.org/10.1109/MIS.2016.29).
- [57] F. Wei and U. T. Nguyen, "Twitter bot detection using neural networks and linguistic embeddings," *IEEE Open J. Comput. Soc.*, vol. 4, pp. 218–230, 2023, doi: [10.1109/OJCS.2023.3302286](https://doi.org/10.1109/OJCS.2023.3302286).
- [58] F. Wei and U. T. Nguyen, "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings," in *Proc. 1st IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Dec. 2019, pp. 101–109, doi: [10.1109/TPS-ISA48467.2019.00021](https://doi.org/10.1109/TPS-ISA48467.2019.00021).



**DANISH JAVED** received the bachelor's and master's degrees in software engineering from Bahria University, Islamabad, Pakistan. He is currently a Ph.D. Scholar with the School of Computer Science, Taylor's University, Malaysia. His work is centered around utilizing advanced computational techniques to extract valuable insights from large datasets and enhancing the transparency and interpretability of AI models. He has contributed to various interdisciplinary projects aimed at enhancing the understanding and application of artificial intelligence in real-world scenarios. His work aims to advance the understanding of social networks and enhance the effectiveness of data-driven decision-making processes. His research interests include text mining, social network analytics, explainable AI, and big data analytics.



**NOOR ZAMAN JHANJHI** (Member, IEEE) is a highly esteemed Senior Professor of computer science, specializing in artificial intelligence and cybersecurity. He currently holds the position of a Professor with the School of Computer Science, Taylor's University, Malaysia, and the Program Director of Postgraduate Research Degree Program and the Director of the Research Center. With a career marked by academic leadership and groundbreaking research, he has been pivotal in advancing research and education in computer science. He recognized globally, he has been ranked among the world's top 2% research scientists for three consecutive years, in 2022, 2023, and 2024. In Malaysia, he is ranked among the top three computer science researchers. His impressive academic portfolio includes more than 70 research books edited or authored with prestigious publishers, such as Springer, Elsevier, Taylor and Francis, Wiley, and IGI Global USA. His contributions to research and innovation are further demonstrated by his successful management of more than 40 internationally funded research grants. His research has garnered more than 1000 points in impact factor, reflecting the depth and breadth of his contributions to the field. A dedicated mentor, he has supervised and co-supervised 38 master's students to successful graduation and has served as an external examiner for more than 60 Ph.D. and master's theses globally.

He was honored with the Outstanding Faculty Member Award by MDEC Malaysia, in 2022, and the Vice Chancellor's Best Research Citations Award from Taylor's University, in 2023. A sought-after keynote speaker, he has delivered more than 70 invited talks at international conferences and has chaired numerous sessions. His commitment to academic excellence and accreditation extends to his involvement with ABET, NCAAA, and NCEAC over the past decade. In addition to his academic accomplishments, he holds an Associate Editor and the editorial board positions with several high-ranking journals, including *PeerJ Computer Science*, *CMC Computers, Materials and Continua*, and *Frontiers in Communications and Networks*. He was also recognized as the Outstanding Associate Editor of IEEE Access.



**NAVID ALI KHAN** received the Ph.D. degree from Taylor's University, Malaysia, with the support of a prestigious full scholarship, where he conducted pioneering research that significantly advanced the field of Remotely Piloted Aircraft technology. He is currently a Senior Lecturer and the Program Director of the School of Computer Science, Taylor's University, where he also mentors and supervises master's and Ph.D. students in cutting-edge research. His work has been published in

several high-impact journals and conferences, contributing significantly to advancements in these fields. With a strong foundation in both academia and industry, he continues to drive innovation in software engineering and UAV technology, aiming to create impactful solutions that bridge theory and practice. His research focuses on UAVs, autonomous systems, and AI-driven automation.



**SAYAN KUMAR RAY** received the Bachelor of Engineering degree from Gulbarga University, India, the Master of Technology degree from the University of Calcutta, India, and the Doctor of Philosophy degree in computer science from the University of Canterbury, New Zealand. He is an accomplished Associate Professor and the Head of School at the School of Computer Science within the Faculty of Innovation and Technology, Taylor's University. In his extensive career, he

has undertaken various academic and administrative roles, showcasing his commitment to education and research. His research spans various areas, including intelligent cyber defense systems, enterprise software application analysis, dynamic spectrum management in 5G, device-to-device communication for post-disaster recovery, and much more. These publications reflect his dedication to advancing technology and addressing crucial issues in the field. Some of his administrative duties include serving as the Chair of the Manukau and Tech Park Campus Research Committee, Academic Leader of Curriculum Development at the School of Digital Technologies, and a member of several important committees and advisory groups at Manukau Institute of Technology, Auckland, New Zealand. His professional memberships further demonstrate his dedication to the field, as he is an active member of numerous Institute of Electrical and Electronics Engineers (IEEE) societies and committees, focusing on topics such as wearable biomedical sensors, connected vehicles, security and privacy, and translational engineering. Furthermore, he has made significant contributions to the field through his publications, with a range of book chapters and articles in journals and conference proceedings.



**ARAFAT AL-DHAQM** received the B.Sc. degree in computer science from the University of Technology, Iraq, and the M.Sc. degree in information security and the Ph.D. degree in computer science from the University Technology Malaysia (UTM). He currently holds a Senior Lecturer position at the School of Computer Science, Taylor's University, Malaysia. He has a solid foundation in information security, digital forensics, information security governance, and risk management. Furthermore, he was trained by Cybersecurity Malaysia (CSM) as a Certified Digital Forensic Investigator and a Certified Information Security Awareness Manager (CISM).



**VICTOR R. KEBANDE** (Member, IEEE) received the Ph.D. degree in computer science (information and computer security architectures and digital forensics) from the University of Pretoria. He was a Researcher with the Information and Computer Security Architectures (ICSA) and the DigiForS Research Groups, University of Pretoria, and he was a Postdoctoral Researcher with the Internet of Things and People (IOTAP) Center, Department of Computer Science, Malmö University, Malmö,

Sweden. He was also a Postdoctoral Researcher of cyber and information security in information systems research subject with the Department of Computer Science, Electrical, and Space Engineering, Luleå University of Technology, Luleå, Sweden. He was a Visiting Researcher with Colorado State University (CSU), USA. He is currently a Researcher with the University of Colorado Boulder, USA, and an Assistant Professor of IT security with the Department of Computer Science (DIDA), Blekinge Institute of Technology (BTH), Karlskrona, Sweden. His research interests include cyber, information security, and digital forensics in the IoT, the IoT security, digital forensics incident response, cyber-physical system protection, critical infrastructure protection, cloud computing security, computer systems, distributed system security, threat hunting and modeling, cyber-security risk assessment, blockchain technologies, and privacy-preserving techniques. He is also an Editorial Board Member of *Forensic Science International: Reports* journal.

• • •