ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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

N.Savitha[1], V.Krishnaveni[2], N.Kamala Vikasini[3]

[1]Assistant Professor, CSE(DS), Swarna Bharathi institute of science and technology, Khammam, TG, India. Email: savitha.natuva@gmail.com

[2]Assistant Professor , CSE, Swarna Bharathi institute of science and technology, Khammam, TG, India. Email: veni.ch1801@gmail.com

[3]Assistant Professor, CSE(DS), Swarna Bharathi institute of science and technology, Khammam, TG, India. Email: vikasini574@gmail.com

ABSTRACT:

Social networking systems such as X (Twitter) function as centers for open human interaction; nonetheless, they are progressively permeated by automated accounts impersonating real users.

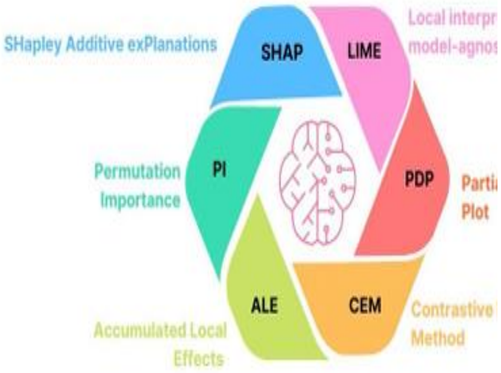
These bots often participate in disseminating misinformation and influencing public sentiment at politically critical periods, such as elections. Many contemporary bot detection techniques depend on black-box algorithms, which raises issues about their transparency and practical applicability. This work seeks to overcome these constraints by formulating an innovative way for identifying spambots and counterfeit followers via annotated data.We present an interpretable machine learning (ML) framework that utilizes various ML algorithms with hyperparameters tuned by cross-validation to improve the detection process.Additionally, we examine several attributes and provide a distinctive feature set targeted for superior performance in bot identification.Furthermore, we use many interpretable AI methodologies, including Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP will elucidate the impact of certain attributes on the model's predictions, aiding in the discernment of whether an account is a bot or a real person. LIME will facilitate understanding of the model's predictions, providing insight into the characteristics that influence the classification outcome. LIME enables researchers to identify bot-like behavior in social networks by producing locally accurate explanations for each prediction. Our approach provides superior interpretability by distinctly illustrating the influence of several variables used for spam and false follower identification, in contrast to current leading social network bot detection techniques. The findings demonstrate the model's capacity to discern critical differentiating features between bots and authentic individuals, providing a clear and efficient solution for social network bot identification.

Furthermore, we use two extensive datasets, Cresci-15 and Cresci-17, which provide solid baselines for comparison. Our approach demonstrates its efficacy by surpassing other techniques while offering interpretability, hence enhancing performance and reliability in bot identification tasks.INDEX TERMS: Interpretable artificial intelligence, social networks, bot identification, fraudulent followers, spam bots.

I. INTRODUCTION:

Social networks have emerged as the primary source of information in the contemporary day. X, formerly referred to as Twitter, is now one of the most prominent and extensively used social media platforms, hence playing a significant part in online discourse and facilitating connections among millions of active users [1]. Nonetheless, its significant social and economic impact has become it a desirable target for malevolent actors. those attempting to control and influence public opinion and decision-making. X has long been a primary target for automated programs, or "bots," because to its open structure and growing user demographic. These bots might be beneficial since authentic bots generate several instructional tweets, including blogs and news updates. Malicious bots, however, propagate spam or detrimental content. The attributes used by contemporary Twitter bot detection algorithms are often based on user data, including timestamps, social connections, behavioral patterns, and network affiliations [2], [3]. Nonetheless, feature engineering takes considerable labor and diligence. Social bots has the capacity to propagate disinformation, including false news, rumors, and hate speech, by swiftly promoting low-credibility material on X via engagements with prominent users and deliberate mentions [4]. The majority of the above listed concerns are managed by the use of bots. A botnet is an assemblage of bots intended to do designated activities, while a Sybil account is a contrived identity that does not correlate with or originate from an actual human user. Botnets and Sybil accounts are often used to spread misinformation and disrupt authentic conversation, hence exacerbating the difficulties of preserving integrity in online forums.

FIGURE 1. Interpretable AI techniques.



A wide variety of domains have found useful applications for machine learning (ML), including sports analytics[7], sentiment analysis[8,9], fake news detection[10], and social bot detection [11]. Interpretable machine learning (XAI) is the subject of our research because of its widespread application in several fields for the purpose of enhancing performance and gaining a deeper understanding of the model. Most often used interpretable AI approaches are shown in Figure 1, with SHAP and LIME being the most popular. Factor analysis, Shapley additive interpretation (SHAP), and local interpretation model-agnostic interpretation (LIME) are a few of the interpretable ML methods that shed light on how a specific data point impacts the prediction model [12]. Stakeholders are able to detect biases in AI systems thanks to the increased transparency, which promotes accountability and equity in AI applications and helps users comprehend and trust these systems. With descriptive ML's help, we can bridge the gap between AI algorithms and human understanding, which in turn allows for more well-informed decisions and more faith in AI. So, to learn more about how social network bot detection (SNBD) works, using XAI is a crucial step [11].

The current body of knowledge distinguishes between real users and automated accounts by analyzing several aspects of the social network. For example, these characteristics may include user activity patterns (such as the number of tweets sent and the timestamps), account information (such as the number of followers to total followers and the average age of the account) and social network structures (such as the number of retweets and mentions) [13], [14], etc.

For this reason, supervised ML models and deep neural networks have been widely used [15], [16]. Heuristic and other traditional bot detection systems aren't up to the task of keeping up with spambots' ever-changing tactics, network-based approaches rely on small social networks, and older ML models ignore patterns in language, time, and sentiment since they use limited features. In addition, most of them cannot be explained, which makes it hard to assess the results. By including several feature sets, our interpretable AI-based model fills these gaps. Through the utilization of XAI, we are able to increase transparency, leading to better accuracy, robustness,and interpretability

In addition, unsupervised detection of aberrant behaviors associated with bots has been investigated using clustering and anomaly detection approaches [17]. Despite the encouraging results, these methods are not always scalable or adaptable because they rely on static datasets and manual feature engineering. Furthermore, there are obstacles to comprehending the decision-making process because to the substantial dependence on black-box ML models, which restricts their interpretability.

Present bot detection approaches are not as effective as they may be due to a number of issues.Feature engineering is one of these obstacles; it's a time-consuming procedure that calls for domain knowledge and human intervention to update the models for use with more recent datasets and bots. Also, bots are able to evolve their techniques to better imitate human users and elude detection algorithms, which exhibits dynamic and adaptive behavior [11]. Consequently, due to the dynamic nature of bot operations, black-box detection methods find it challenging to adjust. Furthermore, these strategies damage confidence and transparency because models are not interpretable. It is difficult to conduct an evaluation without the capacity to understand the results, since this would mean that we have no idea if the model is correctly detecting bots due to significant trends or is just overfitting the data. On top of that, the majority of approaches focus on improving detection accuracy rather than the more generalizable and adaptable aims that are essential for actual social network implementation. More open and interpretable detection frameworks are required to fill these gaps. Thus, the suggested approach takes these difficulties into account by incorporating XAI approaches into bot identification. By illuminating how specific attributes contribute to model predictions, these strategies increase the openness of ML approaches [18].

In light of this, our research provides the following contributions. The area of social network bot detection can benefit from these efforts.

• This paper introduces a novel interpretable bot detection model developed for the purpose of identifying spambots and false followers on Twitter/X and other social media platforms. The model enhances detection by providing clear and interpretable insights on bot identification through the use of interpretable AI approaches.the reliability of the mechanism.

• This research examines all the aspects of X and examines how they impact the model for detecting bots.

In order to improve the model's generalizability, we evaluate it across different types of bots using well-established datasets and use many explainable AI methodologies to examine the behavior of various parameters in the context of bot identification. This study verifies XAI's effectiveness by showing that it outperforms other state-of-the-art algorithms in bot detection with more transparency. In addition to providing insights into the model, the suggested model achieves higher detection outcomes.

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

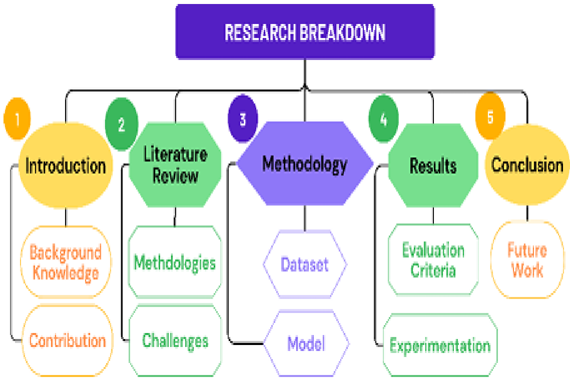


FIGURE 2. Research breakdown.

II. LITERATURE REVIEW:

A deluge of studies offering various techniques have resulted from the academic community's feverish pursuit of answers to bot-detection problems. Regardless, there is a lack of clarity and interpretability in the results provided by most of these methods, thereby leaving a hole in the current literature. Following this, we will give a brief overview of the most popular bot detection algorithms and then look at the problems that need to be solved. Supervised ML paradigms are the backbone of bot detection approaches; to train ML classifiers and build an effective framework, these paradigms require either a single or several annotated datasets. While human annotation is the most common method for creating these annotated datasets, other methods including crowdsourcing, automated annotation techniques, or utilizing pre-existing established models have also been used to build datasets for bot identification. Table 1 provides a summary of the most important works on interpretable AI-based bot identification. Each study's goals and our findings are detailed in this literature.

A. SNBD METHODOLOGIES:

Researchers came up with a unique strategy in [23] by building a database of fake accounts meant to entice spammers and then recording 52248 pieces of information from those accounts' profiles. To make this dataset even more thorough, we added a set of normal user profiles. This allowed us to create a classification algorithm that takes into account both user-centric and content-centric aspects. A different study [24] used a comparable approach to try to identify botnets controlled by the same individual.

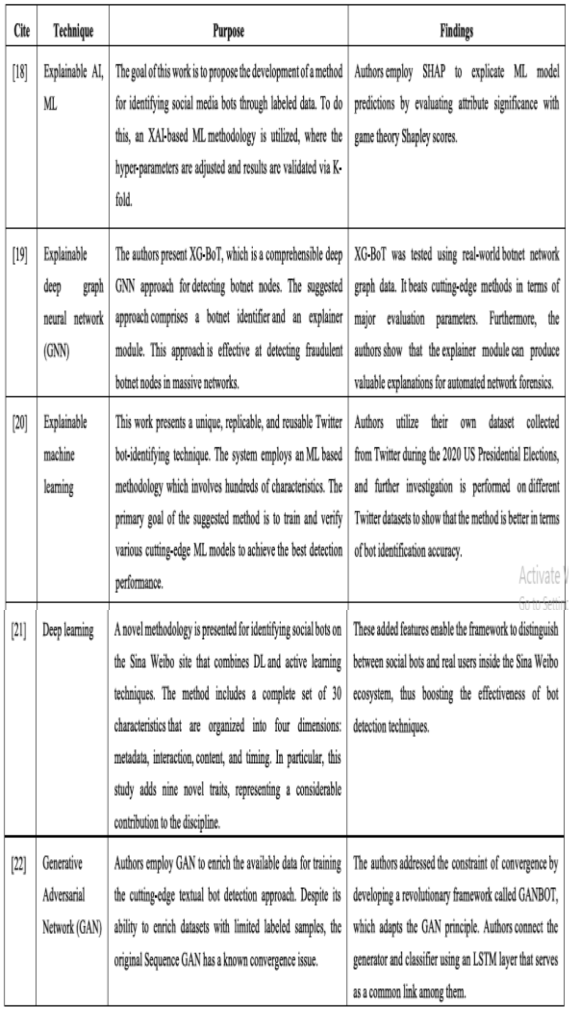
Referring to Reference [25], which uses crowdsourcing techniques for bot recognition on Facebook, the method seemed to work at first. But, as the number of bots continued to grow and evolve, the method became increasingly unscalable, highlighting the need for more adaptive and dynamic bot detection strategies.

Down below, we'll talk about how crowdsourcing has been used for data annotation tasks in various ways. The most popular method, BotOrNot, and its successor Botometer, use a dataset presented by [23] that has been updated with new tweets for each identified account. Its revolutionary feature was the enormous amount of unique attributes used to train the model.

The authors in [46] laid forth a plan for mining this massive feature set and then used a newly annotated dataset to back up their claims. Results confirmed the effectiveness of the proposed paradigm while also drawing attention to certain shortcomings. Since the model was trained on older bot variations with different traits and patterns of activity, its performance suffered when applied to the current dataset. The authors show how to retrain the model and adapt to the changing bot scene by utilizing the crowdfunding features of the Botometer platform. They also provide access to several labeled bot datasets, as mentioned in [27]. Alternatively, Stweeler[28,29] is a relatively simple yet effective method for bot identification that uses a click-bait strategy to collect data on users and tweets. Another method [30] finds bot accounts by looking at how unpredictable the screen name is, while another [31] shows that the trained model is still quite effective even when 10 criteria are carefully selected. Although some methods use Deep Learning (DL) or more complex algorithms, most of the publications that have been discussed use simple ML algorithms.

An example of a DL-based approach to bot detection employing a behavior-augmented model on users is presented in reference [32], which makes use of neural networks. The authors of [23] propose a similar approach, recommending an LSTM network that uses X's content and metadata in conjunction with contextual user attributes to identify automated accounts in tweets. On the other hand, [33] puts up an alternative approach to bot identification, stressing the importance of seeing coordinated assaults rather than individual users. Even though they use a wide variety of features and ML techniques, the tactics discussed above don't seem to successfully handle other problems, which are detailed in the part that follows. The SNBD approach that is most often employed is illustrated in Figure 3.

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:

The aforementioned research show that there have been several attempts to find ways to identify online social bots, yet there are still many unanswered questions. The question of whether or not adding more features improves model efficiency persists, despite the fact that several SNBD methods use more than 1,000 attributes to train their technique [26].

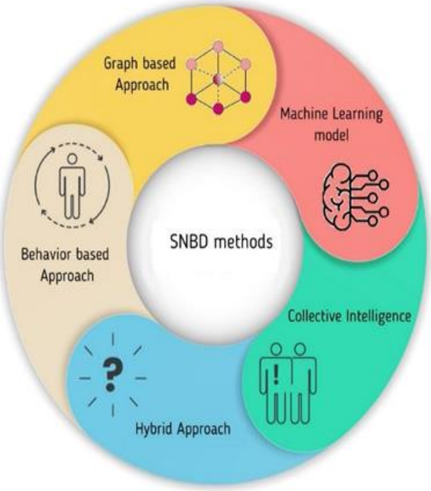
In addition, the authors of the[34]state that using a large feature set has a major effect on how scalable bot detection algorithms are. Notably, the same study also found that using different subsets of publically available labeled datasets can improve the generalizability of models. Notably, bot detection techniques that rely on machine learning have performance that differs across various datasets. Consequently, we need to gather more datasets to make sure our training data covers all the bases in terms of bot behavior. Both [27] and [35] get the

ISSN:0976-0172

same result by providing separate datasets for each type of bot and going to a finer-grained classification of bots. Therefore, defining what characteristics truly make a social bot is a big challenge in online social bot identification.

Malicious purposes involving X bots include spreading misinformation, astroturfing, and fake news [36], [37]. Between the US presidential election of 2016 and the 2018 midterm elections, the authors of [38] looked over 245,000 X profiles and found approximately 31,000 bots. As part of the U.S. Congress's investigation into Russian meddling in the 2016 U.S. election campaigns, the writers of [39] combed over 43 million tweets mentioning the election. According to their research, a sizeable percentage of users, 4.9% of liberals and 6.2% of conservatives, used automated accounts. Importantly, they were able to get recall and precision ratings higher than 90% using their method. The writers of [40] provide an analysis of German

FIGURE 3. SNBD methods.



As shown in the aforementioned research, there are still numerous outstanding challenges in detecting online social bots, despite the proliferation of scientific efforts that have produced diverse methodologies. While it is true that several SNBD methods use over a thousand features for training purposes, the question of whether or not this actually improves model efficiency is still open. Additionally, the authors of the[34]mention how using a large feature set greatly affects the scalability of bot detection systems. Curiously, one thing that was found in the same study is that using different subsets of publically available labeled datasets can improve the generalizability of the models. Machine learning bot detection methods' efficacy differs across datasets. Because of this, we must continue to gather more datasets so that our training data covers all possible bot behavior traits. Results from[27] and[35], which provide separate datasets for each type of bot and proceed to a finer-grained classification of bots, reach the same conclusion. Accordingly, defining what characteristics truly make a social bot is a big challenge in online social bot identification.X bots are frequently employed for malicious purposes, such as spreading false information, orchestrating propaganda, and engaging in astroturfing [36], [37]. From the 2016 US presidential election to the 2018 midterm elections, the authors of [38] tracked 245,000 X profiles and identified approximately 31,000 bots. In order to delve into the matter of Russian influence in the 2016 US election campaigns, the writers of [39] combed through 43 million tweets that were relevant to the investigation by the US Congress. According to their research, a sizeable percentage of users, roughly 4.9% of liberals and 6.2% of conservatives, used automated accounts. Specifically, they were able to get recall and precision ratings higher than 90% using their method. An analysis of German political parties' tweets prior to and during the 2017 election cycle was presented in [40], which shows that the employment of social bots increased. Bot identification on Twitter is obviously not an easy task and often requires strong and comprehensive treatment. A number of One example of an ML-based approach is BotOrNot [26], which offers a grand total of 1200 unique attributes trained using an ML classifier. An improved version of this system called Botometer is described in [27]. However, in order to access user data during real-time computations, it requires X API keys, which makes using real-time labeling tools impractical for big datasets. The Stweeler[28], the Debot[41], and the Retweet-Buster (RTbust) [42] are just a few examples of the growing number of Twitter bot detection systems that employ data statistics and machine learning.

III. METHODOLOGY:

To provide a comprehensive approach for identifying social media spambots and fake followers, our methodology makes use of interpretable AI-based machine learning; this guarantees robustness, generalizability, and interpretability. The first step in building our model, which is based on a modular approach, is to preprocess the dataset with relevant data. After that, we go on to feature engineering, where we choose the optimal qualities for bot identification after extracting many features. User profile features, language features, engagement features, and content-based characteristics are among the many kinds of attributes that we use. Also, we extract sentiment features from textual information like tweets and description text by doing sentiment analysis. The next thing to do is to divide the dataset into three parts: training, validation, and testing.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Make sure that every train-test split uses stratification to keep the class ratio for testing and training data consistent. We employ a number of cutting-edge ML algorithms and explainable AI techniques to conduct thorough testing comparing bot and human classification accuracy. Our goal is to provide a bot recognition solution that is based on machine learning and reliable and accurate. We test the per formance of our model through a varied variety of ML-based algorithmsanduseK-foldcross-validationforresultstoavoid any bias in the model. To ensure a fair comparison, it is essential to apply optimal parameters to find the best version of the classifier, since each machine learning approach has its own unique set of parameters. Our Interpretable AI-based approach, which follows a module-based architecture, is shown in Figure 4. To improve the process of detecting spambots and fraudulent followers, each module carries out a distinct function. The parts that follow provide more information on the methodology.

A. DATASET:

Introduced by [43], the Cresci-15 is an excellent benchmark dataset for detecting bots on social networks. Its purpose is to test how well bot identification algorithms work; it contains both real and fake profiles retrieved from Twitter.

Table 2 shows that the dataset is composed of multiple subsets, each of which represents a different type of bot and human behavior.

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

B.DATA PREPROCESSING:

When it comes to Twitter and other social media platforms, the Cresci-17 dataset is considered the gold standard for bot detection [35]. Unlike any other dataset out there, this one contains tweets from a wide range of accounts, including those of real humans as well as more advanced social bots created and operated with the express purpose of fooling others. The scientific community heavily relies on the Cresci-17 dataset for a wide range of bot detection tasks, including method development, validation, accuracy, generalizability evaluation, and performance comparison.Because of its availability, the area of social network bot detection and mitigation has advanced significantly, leading to the development of more effective strategies. The dataset's attributes are laid out in Table 3.**FIGURE 4.** Proposed Model for identification of spambots and fake followers.

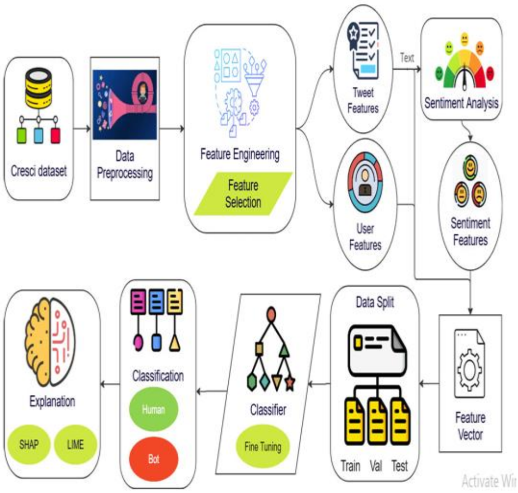


TABLE 3. Dataset characteristics (Cresci-17).

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

The features used in our model were culled from user profiles and tweets. Figure 5 displays the most frequently used keywords in the description text of actual users, whereas Figure 6 displays the same data for bot users, providing insight into the content variety. We can observe that it has many meaningless terms and that some of the words in both clusters are very similar. Hence, it's critical to preprocess the textual input such that the model can differentiate between manually entered text and text generated by bots. The description feature is prepared for by executing a series of preprocessing procedures inside the feature engineering pipeline. This feature is specified by textual information. You can't do sentiment analysis without the description and tweet text data; they let us extract features based on sentiment. Because tree-based classifiers like Random Forest are able to accommodate null values, it is crucial to handle null values inside the data since they can cause complications with these classifiers. Since the description field of X account could have null values, the default value for missing value inferences is "missing," which means that there is no data accessible. But for empty descriptions, the description_length variable remains set to 0. Raw Data from Twitter frequently includes extraneous symbols, URLs, mentions, and emojis that aren't relevant. Preprocessing is the process of cleaning up text data by erasing or replacing these parts. One example is the use of textual representations of emojis for sentiment analysis. Given that our model relies on URL and punctuation information as features, we normalize or eliminate special characters, whitespace, and punctuation to make sure the dataset is consistent and uniform. However, this is done particularly for sentiment analysis. To reduce the quantity of the vocabulary and give more weight to terms with substantial informational substance, we eliminate stop words, which are frequently used but do not convey much meaning

FEATURE SELECTION AND EXTRACTION:

There are two separate files in the dataset, one for users and one for tweets. As can be seen in Table 4, many features are extracted from these two datasets.

FIGURE 5. Description of real users.

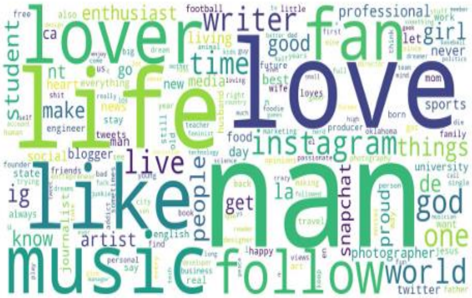
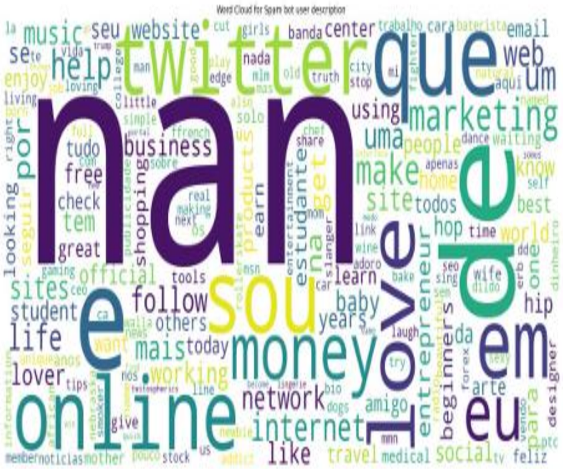
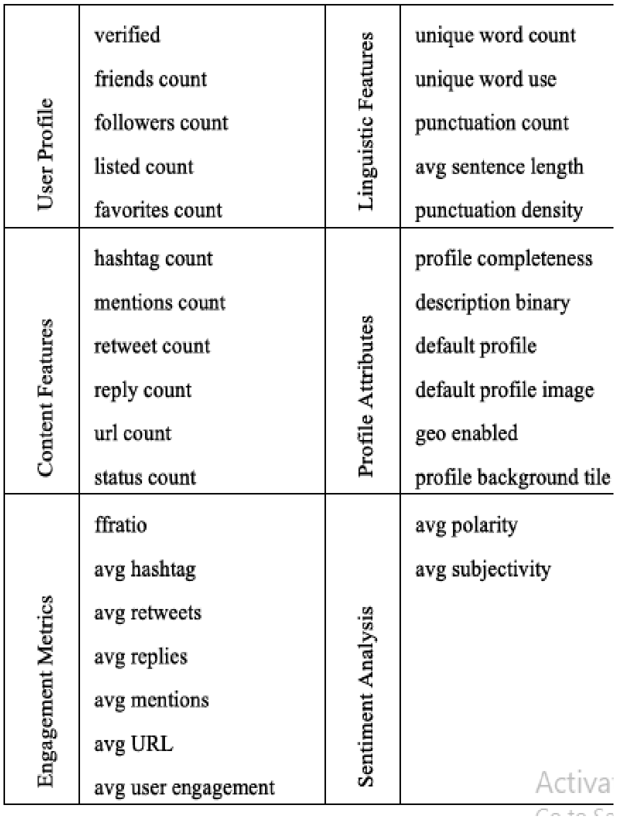


FIGURE 6. Traditional spam bot user description.



Using prior work [3, 45, 46] as a foundation, we studied X's features and created a plethora of features constructed from metadata features. To get the most out of your model, you need to make sure you finish doing feature extraction and selection before you deploy it. Feature selection is essential for enhancing model performance since it determines which traits are the most informative. The input vector's dimensionality is decreased through feature selection, leading to a reduction in method complexity [18]. According to earlier studies, neither a perfect set of qualities nor the ideal quantity of them exist. Differences in datasets affect how well a model trained using a given set of features performs. Unlike other methods, our approach trains a model to differentiate between tweets written by humans and those created by bots using only a small number of tweet- and user-based variables. Retaining good performance while being time efficient is the rationale for employing such a limited set of features. We use extensive datasets that necessitate data preparation, therefore it is crucial to be efficient when selecting features. Utilizing a constrained collection of features is thus essential for expediting computation and retrieval. Our approach is based on Shapley feature selection, which involves utilizing the SHAP method to determine which attributes are most important for differentiating between real and spambot accounts. By assigning a Shapley value to each attribute, which represents its contribution to the prediction, this method clarifies how each element affects the prediction model. By distributing these values over 52252 MLmodels and examples, SHAP may rank the features according on their classification relevance. Different feature groupings in Table 4 offer different ways to look at user behavior on X. For example, "verified" and "friends count" are elements of user profiles that distinguish real accounts from bots that use default settings or have crazy following patterns. While engagement measures show out-of-the-ordinary interaction dynamics, content and linguistic aspects catch inconsistencies in tweet formulation and publishing behavior. To further differentiate bots according to their neutral or programmed tone, we additionally administer sentiment analysis on the textual data, such as text and description, and extract sentiment-based attributes.With its varied feature set, XAI is able to improve model transparency and identify important predictors

TABLE 4. Feature set.



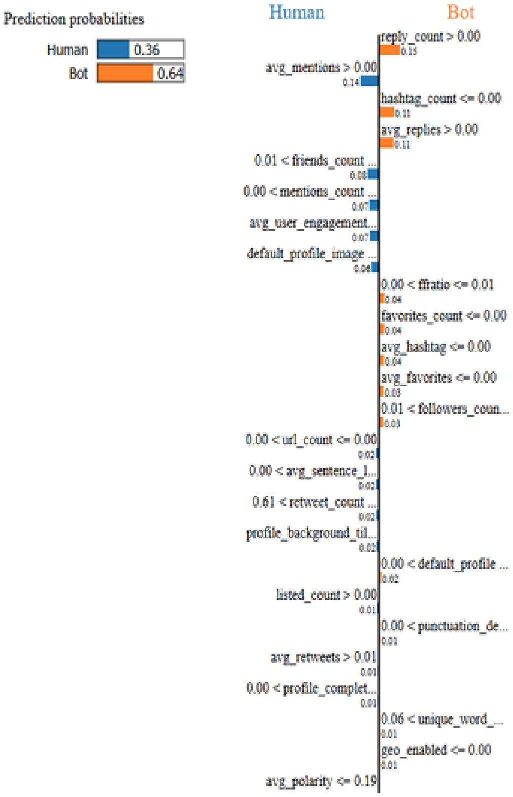
User Profile	verified	Linguistic Features	unique word count
	friends count		unique word use
	followers count		punctuation count
	listed count		avg sentence length
	favorites count		punctuation density
Content Features	hashtag count	Profile Attributes	profile completeness
	mentions count		description binary
	retweet count		default profile
	reply count		default profile image
	url count		geo enabled
	status count		profile background tile
Engagement Metrics	ffratio	Sentiment Analysis	avg polarity
	avg hashtag		avg subjectivity
	avg retweets		
	avg replies		
	avg mentions		
	avg URL		
	avg user engagement		

B. EXPLAINABLE METHOD: LIME:

When it comes to social bot identification in particular, LIME is crucial for making ML models more interpretable. In order to understand how the classifier arrived at its predictions, LIME dissects the role of each attribute in identifying a bot or human account. References [47], [48]. This approach boosts openness and trust by letting researchers examine how various features affect the model's output. The use of LIME to decipher bot detection model predictions is shown in Figures 7 and 8.

The prediction probabilities, as shown on the left side of Figure 7, indicate that the account is probably bot 64% of the time. You can see "Human" and "Bot" split along the middle of the right-hand column. The characteristics that increase the likelihood that an account is a "human" are displayed on the "Human" side of the column. The "bottom" half of the horizontal bar displays the qualities that are indicative of a "Bot" classification. A large number of replies, for instance, is significantly using automated engagement patterns to boost visibility, leading to bot-like behavior. The fact that hashtags are often used by bots to target certain audiences or trends is demonstrated by their 0.11 contribution to the "bot" prediction. A low follower-following ratio, which shows an imbalance in social reciprocity, is also characteristic of bot accounts. However, a retweet count of 0.61 is considered a medium level of retweet value, which is consistent with sharing behavior that is similar to that of a human. A 100% confidence forecast for the word "human" is shown in Figure 8. Examples of favorable indicators for the "Human" classification include numbers that are greater than zero for avg_mentions and favorites_count. However, characteristics that are slightly associated with bot-like activity include default profile and low average hashtag use.

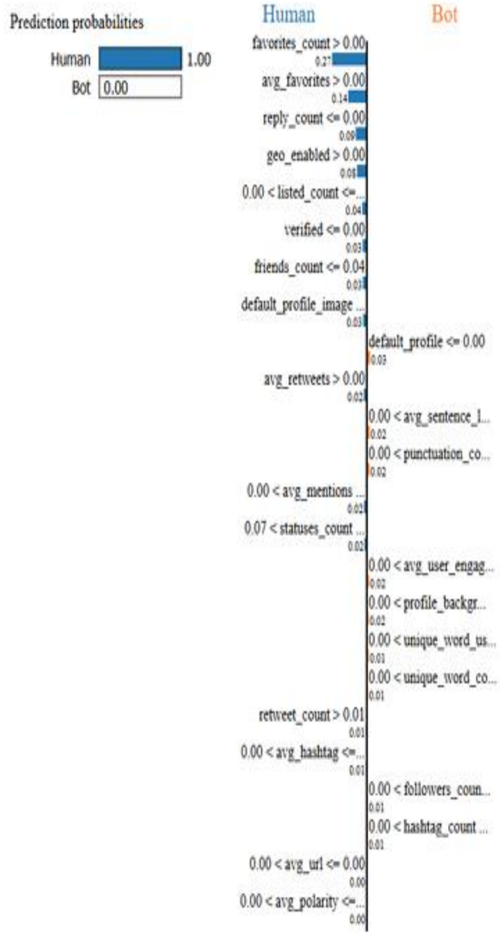
FIGURE 7. Lime explanation for ‘‘bot’’ prediction (cresci-15).



C. EXPLAINABLE METHOD: SHAP

SHAP is a method that sheds light on how specific features contribute to the model's predictions, making ML models for Twitter bot identification more interpretable. SHAP calculates Shapley values, which measure the marginal contribution of each feature to the prediction, by examining all possible combinations of feature subsets. This method has its origins in game theory. By using this approach, the model-independent and open-ended explanation of how features impact the classifier's decision-making can be achieved. We used SHAP to examine and rank the most important features impacting the predictions in our study.The visualization for the cresci-15 dataset is shown in Figure 9, while the SHAP values for the cresci-17 dataset are displayed in Figure 10.

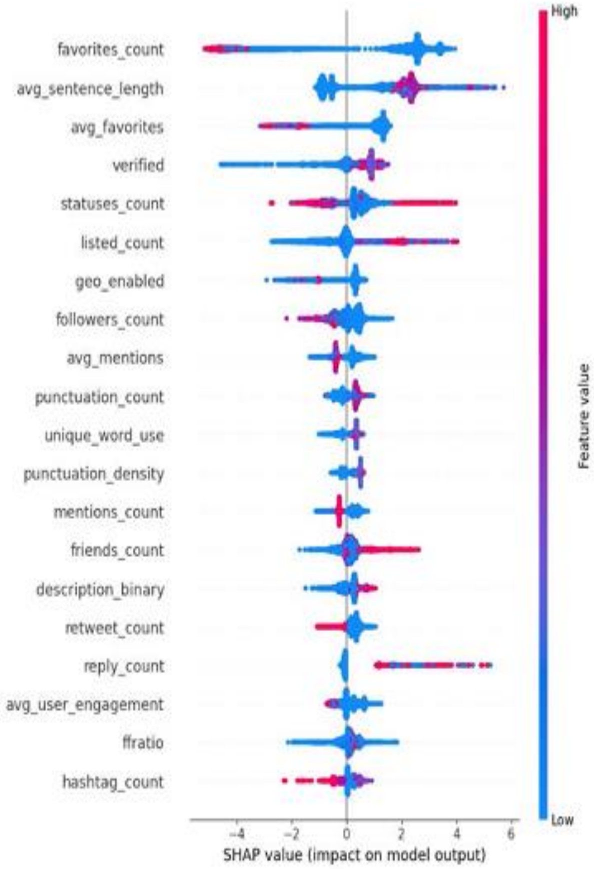
FIGURE 8. Lime explanation for ‘‘human’’ prediction (cresci-17).



The top twenty attributes that significantly impact the ML model's output are shown here. For every feature, one point is given to a specific Twitter user. The real SHAP values, as shown by a point on the x-axis, demonstrate the impact of an attribute on the model's output for that specific X user. If we compare this individual to the average Twitter user, we find that their mathematical likelihood of participating in harmful action is higher. An individual's propensity to engage in hostile behavior on Twitter is inversely proportional to their SHAP score. The significance of features is represented on the y-axis by the average of their absolute Shapley values.

Pictured in Figure 9 are some of the high-value traits that contribute to the prediction, and the SHAP values essentially measure their contribution. Users can be ranked based on their likelihood for malicious activity using this probabilistic interpretation. This enables more targeted interventions to detect and mitigate harmful acts on the site.

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



ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

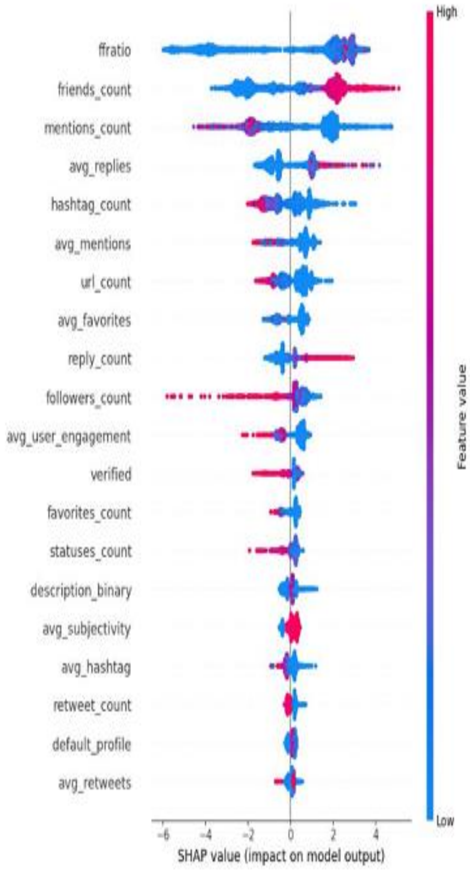
RESULTS: This section assesses the accuracy and generalizability of the suggested SNBDapproach in identifying bots and differentiating them from human users. The investigation delves into the model's capacity to identify bot accounts and evaluates how well it sorts people into normal and bot groups. The first step is'shuffling,' a preprocessing procedure in ML that involves randomly sorting the dataset. This procedure eliminates the possibility of training and testing data that is skewed due to inherent order, such as chronological arrangement or class classification. In addition, we split the data in half, with 75% going into training and 25% into testing.

A. EVALUATION CRITERIA

Several important criteria are used to assess the efficacy of our interpretable ML-based model for detecting spambots and phony followers: The precision: The overall effectiveness of the system's detection is measured by how accurately it classifies an account as authentic or spam-free. F1 metric: One often used metric to evaluate a model's overall performance in classification tasks is the F-measure. As shown in Equation 1, it is determined by summing the two crucial metrics, recall and precision, into a single value. By revealing the percentage of positively identified cases (e.g., 52254 accurately classified bots) out of all instances projected as positively, precision indicates the accuracy of the model's positive predictions. You can figure it out by dividing the total number of positive and negative results by the number of true positives. The equation F1 =2Accuracy + Recall Assessing the interpretability of predictive models built with tools like SHAP and LIME is what interpretability is all about. This statistic assesses how well the model explains its judgments in a way that participants can comprehend, which is crucial for understanding the distribution.

The cresci-17 dataset is explained in Figure 10 (SHAP).Area Beneath the Line: The sensitivity rate, which is the rate of true positives, and the false positive rate are both measured.It can take on values between 0 and 1, with 1 indicating flawless classification, 0.5 indicating random guessing, and values closer to 0 indicating subpar performance. Validation with K-Folds: In the context of employing explainable AI-based ML to detect social media spam bots and phony followers. Both the 70%-30% retention approach and 5-fold cross-validation were used to analyze the data using different ML classifiers. A 5-fold cross-validation procedure involves splitting the dataset into five equal parts; one part serves as the test set, while the other parts serve as the training set.

FIGURE 10. SHAP explanation for cresci-17 dataset.



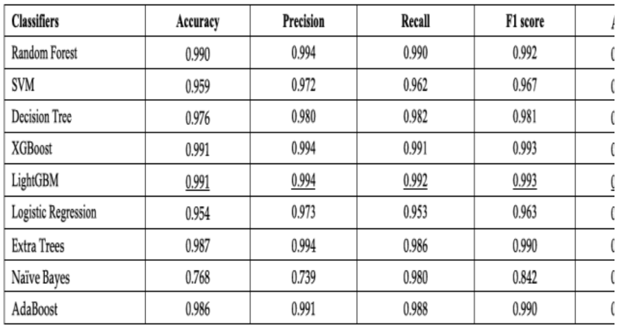
CLASSIFICATION RESULTS The classification results from the two datasets stated in section III-A are discussed in this section. In order to develop a more reliable model for bot detection, we ran multiple machine-learning classifiers on these datasets. The findings precede acquired using K-fold cross-validation with a value of 5 and presented in Tables 5 and 6. As a result, the eliminates the problem of overfitting and offers consistent outcomes. For a more thorough grasp of the outcomes, we display them in relation to recall, accuracy, precision, F1 score, and AUC. Based on the results shown in Table 5, LightGBM outperforms the other classifiers tested on the cresci-15 dataset in terms of accuracy (0.991) and F1 (0.993), but it lags slightly behind in recall (0.093). Table 6 displays the outcomes for the cresci-17 dataset. XGBoost and Light GBM achieved the highest accuracy with F1 scores of 0.990 and 0.993, respectively, out of various classifiers examined. With the exception of Naive Bayes and Support Vector Machines, which significantly decrease accuracy and F1, all classifiers perform competitively.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

To gain a better understanding of how each classifier deals with false positives, it is crucial to examine the trade-offs between recall and precision. In order to prevent the needless blocking of accounts belonging to legitimate users, it is essential to minimize false positives. Results showing our model's excellent accuracy across different kinds of bots and datasets demonstrate that our model minimizes false positives.

TABLE 5. Results on the cresci-15 dataset.



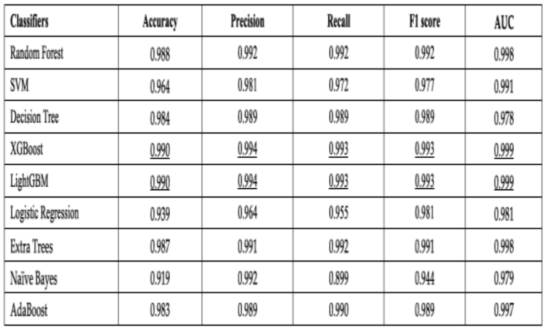
Classifiers	Accuracy	Precision	Recall	F1 score
Random Forest	0.990	0.994	0.990	0.992
SVM	0.959	0.972	0.962	0.967
Decision Tree	0.976	0.980	0.982	0.981
XGBoost	0.991	0.994	0.991	0.993
LightGBM	0.991	0.994	0.992	0.993
Logistic Regression	0.954	0.973	0.953	0.963
Extra Trees	0.987	0.994	0.986	0.990
Naïve Bayes	0.768	0.739	0.980	0.842
AdaBoost	0.986	0.991	0.988	0.990

IV. DISCUSSION AND COMPARISON

In this study, we demonstrate that interpretable ML-based models can successfully detect X-like social network platforms' spam bots and false followers. By creating an interpretable model that makes use of methods like SHAP and LIME, our work aimed to solve the shortcomings of existing bot detection methods. By shedding light on the relevance and interpretation of the features utilized in SNBD, these techniques improve our comprehension of the decision model. For instance, according to SHAP analysis, the prediction model is impacted by high-value indicators like favorites_count, average_mentions, unique_word_use, and followers-following ratio. Because it enables models to be accurate and exact, this transparency is vital for constructing trust in these methods. Several significant challenges in the SNBD job are addressed in this research. Firstly, it explains the contribution of characteristics using XAI and decreases the black-box nature of typical bot identification methods. Due to the high dimensionality and huge sample sizes of datasets like Cresci-15 and Cresci-17, the usage of XAI can lead to computational overhead, which poses a substantial problem. When working with large feature sets, SHAP's computational expense might escalate due to its dependence on estimating shapley values, which creates exponential temporal complexity. In response to this, we give a small feature set that achieves competitive performance with the help of 31 features. Similarly, runtime in scenarios involving enormous volumes of data is increased by LIME's requirement to train local surrogate models for each prediction. However, if we are just interested in a small subset of instance prediction findings, LIME might not be the most expensive option. The use of XAI in large-scale frameworks could be hindered by these computing needs. To fix these problems, optimization techniques like dimensionality reduction can be used to reduce the overall number of features while keeping performance high. It is of the utmost importance that our SNBD model be able to generalize, thus we train it on two massive datasets that contain several types of bot accounts, including social spambots, classic spambots, and phony followers. Our model is trained to reliably and accurately identify different kinds of bots and real users. Models used for bot detection on social networks are not interpretable, which causes a number of restrictions. Because they make forecasts without revealing their decision-making process, black-box models erode trust due to their lack of transparency. Because of this, providing a rationale for classification results is difficult. In addition, finding and fixing misclassifications like false positives or negatives becomes more complicated due to the lack of interpretability, which in turn hampers debugging. Furthermore, without understanding how features contribute, non-interpretable models are more likely to be biased by the training data, which in turn produces discriminatory results.

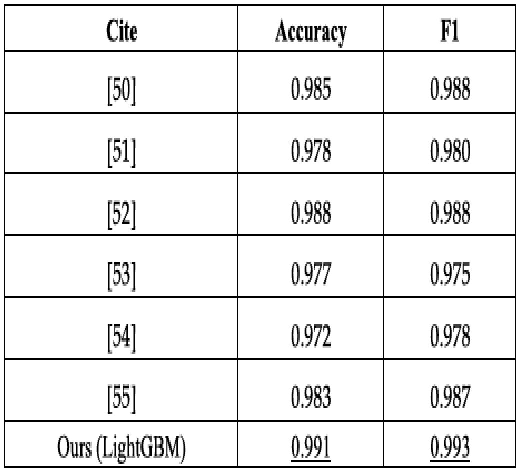
In addition, social media bots are notorious for their ever-changing tactics to avoid detection, and models struggle to keep up with these developments because to their lack of interpretability. Relying on static traits that may become irrelevant with time is a potential outcome of this rigidity. Lastly, chances for improving are limited since non-interpretability prevents us from gaining insights on feature importance. modeling and recognizing important signs of bot behavior. In order to overcome these constraints, XAI techniques significantly increases the efficacy of bot identification on platforms such as X by significantly improving model transparency and adaptability.

TABLE 6. Results for cresci-17 dataset.



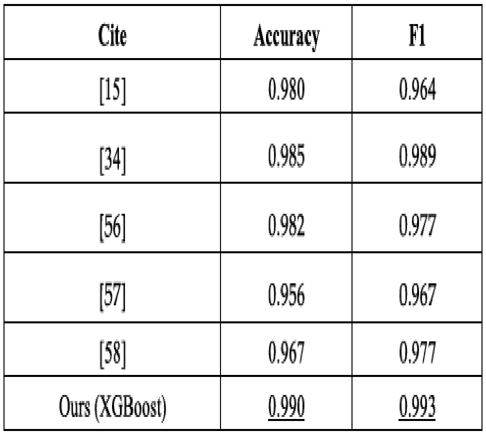
Classifiers	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	0.988	0.992	0.992	0.992	0.998
SVM	0.964	0.981	0.972	0.977	0.991
Decision Tree	0.984	0.989	0.989	0.989	0.978
XGBoost	0.990	0.994	0.993	0.993	0.999
LightGBM	0.990	0.994	0.993	0.993	0.999
Logistic Regression	0.939	0.964	0.955	0.981	0.981
Extra 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)	0.991	0.993

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)	0.990	0.993

Tables 7 and 8 compare results from two research that used the cresci-15 and cresci-17 datasets, respectively. While some of the baseline data were gathered by physical inspection, the vast majority were culled from references [49] and [50]. Our suggested models outperform the competition when it comes to detecting bots on social networks, as shown by the findings. With an accuracy of 0.991 and an F1 score of 0.993, LightGBM had the best performance on the cresci-15 dataset. With an F1-score of 0.993 and an accuracy of 0.990 on the Cresci-17 dataset, the XGBoost model beats all current state-of-the-art algorithms in terms of 52256 F1, recall, and precision.

This demonstrates its exceptional accuracy in detecting bot accounts while also minimizing false positives, demonstrating a remarkable equilibrium between recall and precision. Table 4 shows that our approach makes use of several rich characteristics that capture various aspects of user behavior, linguistic patterns, content qualities, and sentiment analysis. Our feature engineering provides a multi-pronged approach that substantially improves classification performance, in contrast to traditional bot detection models that rely on a handful of network-or profile-based attributes. In addition, using XAI for bot detection aids in choosing the most characteristics that render earlier models ineffective

IV.CONCLUSION

Using an interpretable ML framework that harvests and analyzes features for the purpose of SNBD, this research proposes a new technique to discriminate between bots and actual users on X. A variety of features extracted from the datasets covered in Section III-A make up the suggested methodology. To enhance the model's ability to detect social and spam bots, as well as phony followers, it was trained on a variety of features that were refined using explainable AI methodologies. Our model's accuracy and reliability were both boosted by this method, and we gained valuable insights into possible trends that improved social media transparency. security. Researchers are able to comprehend the effects of the characteristics on the model by integrating the XAI methods SHAP and LIME into the model. With this knowledge in hand, we were able to distill the feature set down to its essentials, relieving the ML model of some of its burden. This study's importance rests in the fact that it provides an interpretable methodology that helps close the gap between model accuracy and transparency, thereby tackling the main obstacles in bot detection.This method allows for more effective bot identification by increasing the trustworthiness of detection models and giving practical insights into the importance of features. Our model is still not perfect because it only uses a limited set of features. Dealing with fresh bots might make this tough. The persistence of new-generation bots makes it all the more difficult to detect their attempts to imitate human behavior. in order to develop increasingly complex behaviors. Additional studies have the opportunity to study adaptive models that can incorporate continuous learning and learn from changing bot behaviors. ways to keep up with these developments. Integrating graph neural networks could improve feature representation and extraction, which is especially relevant in social networks due to their interaction-based nature. To further understand network behaviors and bot identification, future studies should look at how to combine graph-based representations with explainable AI.

CONTRIBUTIONS

Conceptualization, Danish Javed; Formal analysis, Danish Javed, Noor Zaman Jhanjhi; Funding acquisition, Sayan Kumar Ray, Arafat Al-Dhaqm, Victor R. Kebande; Investi gation, 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-Cantallops, 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.

[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.

[4] X.ZhangandA.A.Ghorbani, ''An overview of online fake news: Charac terization, detection, and discussion,'' Inf. Process. Manage., vol. 57, no. 2, Mar. 2020, Art. no. 102025, doi: 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.

[6]Z.Yang,C.Wilson,X.Wang,T.Gao,B.Y.Zhao,andY. 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. [7] D. Javed, N. Z. Jhanjhi, and N. A. Khan, ''Football analytics for goal pre diction 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.

[8]M.Humayun,D.Javed,N.Jhanjhi,M.F.Almufareh,andS.N.Almuayqil, ''Deep learning based sentiment analysis of COVID-19 tweets via resam pling and label analysis,'' Comput. Syst. Sci. Eng., vol. 47, no. 1, pp. 575–591, 2023.