International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

AI-POWERED DETECTION OF FAKE FOLLOWERS AND SOCIAL MEDIA SPAMBOTS USING EXPLAINABLE MACHINE LEARNING

DR. SRINIVAS RAO NIDAMANURU[1], AKARAPUR SATHVIKA[2], G SIDDU RAMESH[3], BULLA NANDAN REDDY[4], ABDUL HUZAIFA[5], BACHALA SHARATH KUMAR[6]

PROFESSOR[1], UG SCHOLAR[2,3,4,5&6]

DEPARTMENT OF CSE, NARSIMHA REDDY ENGINEERING COLLEGE (UGC- AUTONOMOUS) MAISAMMAGUDA (V), KOMPALLY, SECUNDERABAD, TELANGANA-500100

PROBLEM STATEMENT The rapid growth of social media platforms has led to a significant increase in the presence of spambots and fake followers, which are often used to manipulate online influence, spread misinformation, promote malicious content, and artificially increase user popularity. These automated accounts can negatively affect the credibility and trustworthiness of social networks. Traditional detection methods mainly rely on rule-based systems or basic machine learning techniques, which often struggle to identify sophisticated spambots that mimic human behavior. Additionally, many existing AI models operate as black-box systems, making it difficult to understand how decisions are made. This lack of transparency limits trust and practical adoption in real-world applications. Therefore, there is a need for an accurate, transparent, and interpretable AI-based system that can effectively detect fake followers and spambots while providing clear explanations for its predictions

OBJECTIVES

The main objectives of the proposed system are:

1. To develop an AI-based framework capable of detecting fake followers and spambots on social media platforms.
2. To apply machine learning algorithms that analyze user behavior, profile features, and activity patterns to identify suspicious accounts.
3. To incorporate Explainable AI (XAI) techniques that provide interpretable insights into the decision-making process of the model.
4. To improve the accuracy and reliability of spambot detection compared to traditional methods.
5. To enhance transparency and trust in automated detection systems used by social media platforms.
6. To assist social media administrators in maintaining platform integrity by identifying and removing malicious accounts.

ABSTRACT

The increasing use of social media platforms has resulted in the widespread presence of spambots and fake followers, which can manipulate online interactions, spread misinformation, and reduce the credibility of digital platforms. Detecting such malicious accounts has become a critical challenge due to their ability to imitate human behavior and evade traditional detection mechanisms. This study proposes an **AI-powered system for detecting fake followers and social media spambots using explainable machine learning techniques**. The system analyzes various user attributes, including account metadata, activity patterns, follower relationships, and posting behavior, to identify suspicious accounts. Machine learning algorithms are employed to classify accounts as genuine or malicious, while Explainable AI methods are integrated to provide transparent explanations for model predictions. This approach not only improves detection accuracy but also enhances interpretability and trust in the system. The proposed framework can assist social media platforms in maintaining authentic user engagement, preventing manipulation of online influence, and improving overall platform security.

1.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 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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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. 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 botdetection 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 mode

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

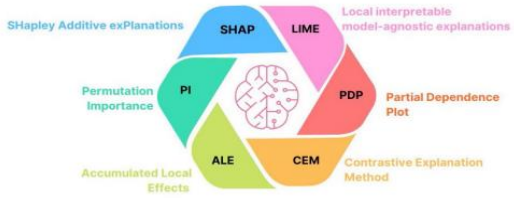

FIGURE 1. **Interpretable AI techniques.**

2.LITERATURE REVIEW

Social media platforms such as Twitter (X), Facebook, and Instagram have become essential communication channels for individuals, organizations, and governments. However, these platforms are increasingly affected by **fake followers, spambots, and automated accounts** that manipulate information, spread spam, and artificially increase popularity metrics. Detecting such malicious accounts has become an important research area, leading to the development of **Artificial Intelligence (AI) and Machine Learning (ML)–based detection systems**. This literature review summarizes key research contributions, methodologies, and challenges related to AI-based fake follower and social media spam bot detection.

Early Research on Social Media Spammer and Fake Account Detection

Early studies focused on identifying suspicious accounts through **behavioral analysis and rule-based approaches**. Researchers analyzed characteristics such as posting frequency, follower-following ratio, and abnormal interaction patterns.

One of the pioneering works by **Stringhini et al. (2010)** used machine learning with social honeypots to identify spammers on social networks. The study demonstrated that features such as **profile attributes, message content, and network interactions** can effectively distinguish malicious accounts from genuine users. These methods laid the foundation for automated detection techniques in online social networks.

Later studies showed that fake followers are often created to **manipulate popularity and influence** on social platforms. Cresci et al. (2015) analyzed fake follower markets and developed classifiers capable of identifying more than **95% of fake follower accounts** using behavioral and profile-based features.

Machine Learning Approaches for Bot and Fake Follower Detection

With the growth of social media data, machine learning techniques became widely adopted for bot detection. Researchers applied supervised algorithms such as **Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes, and Logistic Regression** to classify accounts as human or bot.

Machine learning models typically rely on features extracted from user profiles, including:

- Account age
- Posting frequency
- Follower-following ratio
- Content similarity
- Hashtag usage and engagement patterns

These features allow ML algorithms to detect suspicious patterns that indicate automated or fraudulent behavior. Studies show that ML models significantly improve detection accuracy compared with manual or rule-based methods.

A comprehensive review of machine learning-based social media bot detection highlighted that both **supervised, semi-supervised, and unsupervised techniques** are used to classify accounts across platforms such as Twitter, Facebook, Instagram, and LinkedIn.

Deep Learning Techniques for Social Media Bot Detection

Recent research has explored **deep learning approaches** to handle large-scale and complex social media datasets. Models such as **Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM)** are used to analyze textual content, temporal patterns, and network structures.

A systematic review of deep learning methods indicates that these models can effectively differentiate between human and automated accounts by analyzing **high-dimensional features and temporal behavior patterns**. Deep learning techniques often achieve higher accuracy compared with traditional machine learning approaches.

Additionally, hybrid frameworks combining **deep learning and graph-based analysis** have been proposed to detect coordinated bot campaigns and fake follower networks.

Explainable Artificial Intelligence (XAI) in Bot Detection

Although AI models provide high detection accuracy, many models function as **black-box systems**, making it difficult to understand how predictions are made. To address this issue, researchers have introduced **Explainable Artificial Intelligence (XAI)** techniques to provide transparent decision-making.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Recent studies propose interpretable frameworks that combine machine learning models with explanation techniques such as **LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations)**. These approaches highlight the most influential features contributing to the classification of an account as a bot or human.

Explainable AI improves trust in detection systems by enabling analysts and social media platforms to understand **why an account is flagged as suspicious**, thereby improving model reliability and decision transparency.

Datasets and Feature Engineering for Bot Detection

Many studies emphasize the importance of **high-quality datasets** for training AI models. Commonly used datasets include **Cresci datasets, TwiBot-20, and TwiBot-22**, which contain labeled bot and human accounts for training and evaluation.

The **TwiBot-22 dataset**, introduced in 2022, provides large-scale graph-based data that includes social network structure, user metadata, and behavioral information. Such datasets help overcome limitations of earlier datasets, such as small sample sizes and incomplete annotations.

Feature engineering also plays a crucial role in detection performance. Researchers typically extract three major categories of features:

1. **Profile features** – username, bio, profile picture, account age
2. **Behavioral features** – posting frequency, retweet patterns, interaction rates
3. **Network features** – follower connections, community structure, clustering behavior

Combining these features improves classification accuracy and helps detect coordinated bot networks.

Challenges in Detecting Fake Followers and Spambots

Despite significant progress, several challenges remain in social media bot detection:

- **Evolving bot behavior:** Modern bots mimic human behavior to avoid detection.
- **Class imbalance:** Genuine accounts significantly outnumber fake accounts in datasets.
- **Adversarial techniques:** Bot developers continuously modify patterns to evade detection systems.
- **Lack of interpretability:** Many deep learning models remain difficult to explain.

Researchers emphasize the need for **hybrid AI models, improved datasets, and explainable AI techniques** to address these challenges.

Research Gap and Future Directions

Although numerous machine learning and deep learning techniques have been proposed for bot detection, several gaps still exist:

- Limited integration of **explainable AI with bot detection models**
- Insufficient cross-platform detection systems
- Lack of real-time detection frameworks
- Limited datasets for emerging social media platforms

Future research should focus on **interpretable hybrid AI models**, real-time detection algorithms, and large-scale datasets that incorporate behavioral and network features.

SYSTEM ANALYSIS

EXISTING SYSTEM

- 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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

dynamic bot detection strategies. The most common method [26], BotOrNot and subsequentlyBotometer, 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 crowd funding 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.

DISADVANTAGES

- The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets for Identification of Spam bots and Fake Followers.
- Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.
- Incorrect labeling: The existing machine learning models are only as accurate as the data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

PROPOSED SYSTEM

In 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 ourresearch. 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.

ADVANTAGES

- 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 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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.

IMPLEMENTATION

MODULES

SERVICE PROVIDER

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse and Train & Test Data Sets ,View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Identification of Spambot and Fake Follower Status, Find Identification of Spambot and Fake Follower Ratio, Download Predicted Data Sets, View Identification of Spambot and Fake Follower Ratio Results, View All Remote Users.

VIEW AND AUTHORIZE USERS

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

REMOTE USER

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, Identification of Spambot and Fake Follower.

BOT DETECTION

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.

ALGORITHMS

DECISION TREE CLASSIFIERS

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class
Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

GRADIENT BOOSTING

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-NEAREST NEIGHBORS (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not "learn" until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Example

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

LOGISTIC REGRESSION CLASSIFIERS

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

NAÏVE BAYES

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

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.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[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.

[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, 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.

[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.

[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.

[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.

[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.

[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. Shaahid, 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.

[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.

[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.