Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Cybersecurity in social networks: An ensemble model for Twitter bot detection

Abdulbasit A. Darem [1,*], Asma A. Alhashmi [1], Meshari H. Alanazi [1], Abdullah F. Alanezi [1], Yahia Said [2], Laith A. Darem [2], Maher M. Hussain [3]

[1]*Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia*
[2]*Department of Electrical Engineering, College of Engineering, Northern Border University, Arar, Saudi Arabia*
[3]*Department of Civil Engineering, College of Engineering, Northern Border University, Arar, Saudi Arabia*

ARTICLE INFO

Article history:
Received 18 July 2024
Received in revised form
1 September 2024
Accepted 30 October 2024

Keywords:
Bot detection accuracy
Ensemble learning methods
Social media integrity
Machine learning classifiers
Model interpretability

ABSTRACT

The increasing presence of bot accounts on social media platforms creates major challenges for ensuring truthful and reliable online communication. This study examines how well ensemble learning techniques can identify bot accounts on Twitter. Using a dataset from Kaggle, which provides detailed information about accounts and labels them as either bot or human, we applied and tested several machine learning methods, including logistic regression, decision trees, random forests, XGBoost, support vector machines, and multi-layer perceptrons. The ensemble model, which merges predictions from individual classifiers, achieved the best performance, with 90.22% accuracy and a precision rate of 92.39%, showing strong detection capability with few false positives. Our results emphasize the potential of ensemble learning to improve bot detection by combining the strengths of different classifiers. The study highlights the need for reliable and understandable detection systems to preserve the authenticity of social media, addressing the changing tactics used by bot developers. Future research should explore additional types of data and ways to make models easier to understand, aiming to further improve detection results.

1. Introduction

In today's digital age, social media platforms are essential for communication, information dissemination, and social interaction. However, these platforms face significant challenges due to the proliferation of automated accounts, commonly referred to as bots. Bots are programmed to mimic human behavior and engage with users, often with malicious intent. They can spread misinformation, manipulate public opinion, and commit fraud, thereby undermining the integrity of social media ecosystems (Ferrara et al., 2016; Varol et al., 2017; Shao et al., 2018). Detecting bot accounts on social media is crucial for maintaining the authenticity and trustworthiness of these platforms. Research has shown that a substantial proportion of active users on Twitter are bots, with estimates ranging from 9% to 15% (Varol et al., 2017). This prevalence poses a threat to the credibility of information shared on these platforms and can significantly distort public discourse (Shao et al., 2018). Consequently, there is an urgent need for effective methods to identify and mitigate the impact of bot activity on social media. Various techniques have been developed to detect bot accounts, including rule-based systems, machine learning algorithms, and graph-based approaches. Rule-based systems rely on predefined patterns and heuristics, such as unusual account creation dates or specific keywords, to flag suspicious accounts (Wang, 2010). Machine learning algorithms, on the other hand, leverage labeled datasets to learn distinguishing features of bot accounts and improve detection accuracy (Varol et al., 2017). Graph-based methods analyze the network structure and interactions between accounts to uncover anomalous behavior indicative of bots (Ferrara et al., 2016).

Despite these advancements, detecting bots remains a challenging task due to high false positive rates and the evolving strategies employed by bot creators (Cresci et al., 2017). To address these limitations, this research proposes the use of ensemble learning techniques, which combine

* Corresponding Author.
Email Address: basit.darem@nbu.edu.sa (A. A. Darem)
https://doi.org/10.21833/ijaas.2024.11.014
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-5650-1838

multiple machine learning algorithms to enhance the accuracy and robustness of bot detection on Twitter. Ensemble learning involves training diverse models and aggregating their predictions to achieve more reliable classifications (Dietterich, 2000). By integrating the strengths of individual algorithms and mitigating their weaknesses, ensemble learning can provide a more effective solution for identifying bot accounts. This paper aims to investigate the efficacy of ensemble learning in detecting suspicious Twitter accounts. By leveraging a comprehensive dataset and employing various machine learning models, we seek to develop a robust system that can accurately distinguish between human and bot accounts. The proposed approach not only improves detection accuracy but also contributes to the broader effort of maintaining the integrity of social media platforms. The main contributions of this research are:

- To develop a robust system that can accurately distinguish between human and bot accounts.
- The study also demonstrates the superior performance of ensemble learning techniques in bot detection on Twitter.
- The research underscores the importance of effective feature engineering, such as the followers-to-followings ratio and account age, in enhancing the predictive power of bot detection models.
- The study identifies the ongoing challenges posed by evolving bot strategies and emphasizes the necessity for continuous adaptation of detection methods.

The paper is structured as follows: The Introduction section provides background on the prevalence and impact of bot accounts on social media, particularly Twitter, and introduces the use of ensemble learning for bot detection. The Related Work section reviews existing bot detection methods, including feature-based and machine learning approaches, and discusses ensemble-learning applications. The Methodology section details the dataset, feature engineering, and the machine learning algorithms used, including the ensemble learning approach and evaluation metrics. The Results and Discussion sections present and analyze the performance of individual classifiers and the ensemble model, highlighting strengths, weaknesses, and computational efficiency. Finally, the Conclusion summarizes the findings, discusses implications, and suggests future research directions.

2. Literature survey

The detection of bot accounts has primarily relied on feature-based methods that identify distinctive attributes of bots. Yang et al. (2020) categorized these features into four main groups: User profile, content, temporal, and network features. User profile features include details like account age, screen name, and profile description. Content features focus on the linguistic and semantic aspects of tweets. Temporal features analyze the timing and frequency of activities, while network features examine the relationships and interactions between accounts. Research has delved into various combinations of these features to enhance bot detection accuracy. Kudugunta and Ferrara (2018) proposed an approach that combines account-level and tweet-level features, such as follower count, follower-to-following ratio, and the presence of URLs in tweets. This method demonstrated high accuracy in identifying bot accounts, particularly those involved in political discourse during the 2016 U.S. presidential election. However, feature-based methods face challenges due to the evolving nature of bot behavior, which can render manual feature engineering obsolete (Cresci et al., 2020).

Machine learning algorithms have emerged as powerful tools for bot detection, capable of capturing complex patterns and adaptive bot strategies. Supervised learning algorithms, such as decision trees, random forests, and support vector machines (SVM), are widely used in this domain (Rauchfleisch and Kaiser, 2020). These algorithms are trained on labeled datasets where each account is classified as either a bot or a human. Random forests and decision trees have shown strong performance in bot detection. Minnich et al. (2017) utilized a random forest classifier, achieving high accuracy by leveraging both account-level and content-based features. Similarly, Knauth (2019) applied decision trees to identify bots in a dataset of German political tweets, demonstrating the algorithm's effectiveness in capturing complex decision rules. Support vector machines have also been effective in bot detection tasks. Fernquist et al. (2018) employed an SVM classifier to identify bots spreading false information during the 2016 U.S. presidential election. Their model, which incorporated account creation dates and tweet content, achieved high accuracy. Moe and Schweidel (2017) used SVMs to detect bots in COVID-19-related tweets, highlighting the importance of temporal and content-based features in crisis situations. Deep learning algorithms, particularly convolutional neural networks (CNN) and recurrent neural networks (RNN) have recently advanced the field of bot detection. Kudugunta and Ferrara (2018) successfully used CNNs to detect bots based on tweet content. Yang et al. (2020) introduced a deep learning framework combining CNNs and RNNs to capture spatial and temporal patterns in bot behavior, outperforming traditional machine learning algorithms.

Ensemble learning strategies have gained prominence in bot detection for their ability to enhance classification performance and robustness. Ensemble methods combine multiple models to leverage their diverse strengths and offset individual weaknesses (Sagi and Rokach, 2018). The three main ensemble techniques are bagging, boosting, and stacking. Bagging, or bootstrap aggregating, trains multiple base models on different subsets of training

data and combines their predictions through majority voting or averaging (Sagi and Rokach, 2018). Random forests, an ensemble of decision trees, are a common application of bagging in bot detection (Minnich et al., 2017; Rauchfleisch and Kaiser, 2020). Boosting algorithms, such as AdaBoost and gradient boosting, iteratively train models on weighted versions of the training data, emphasizing misclassified samples. These models combine predictions using a weighted voting strategy. Gradient boosting algorithms like XGBoost and LightGBM have demonstrated high performance in bot detection tasks (Knauth, 2019; Yang et al., 2020). Stacking involves training a set of base models and using their predictions as input for a meta-model, which learns to optimally combine these outputs. Ilias et al. (2024) utilized stacking in bot detection, integrating decision trees, random forests, and deep learning models to achieve superior performance compared to individual models.

In addition to previous studies, recent advancements in 2024 have further pushed the boundaries of bot detection techniques. Bibi et al. (2024) introduced a novel transfer learning-based deep neural network (DNN) model for Twitter bot detection. Their model, TL-PBot, effectively utilized pre-trained models and fine-tuned them on bot detection tasks, significantly improving detection rates on previously unseen datasets.

Ilias et al. (2024) further expanded the application of multimodal transformers in bot detection by integrating text, image, and network data into a unified framework. This approach allowed for a more holistic analysis of bot behavior across different content types, which is particularly relevant in the context of social media platforms that increasingly rely on multimedia content.

Levonian et al. (2021) provided insights into the complex interactions within online communities, which is relevant to bot detection as it underscores the importance of understanding the patterns of engagement between users. Their research, while focused on mutual support connections in an online health community, highlighted the role of nuanced behavioral patterns, which is crucial when distinguishing between human users and automated bots. This study contributes to the broader understanding of social interactions online, offering potential methodologies that could be adapted for bot detection by analyzing engagement patterns that deviate from typical human behavior.

Building on the importance of behavioral analysis, Sallah et al. (2023) explored the use of transformer-based models for detecting bots on Twitter. Their study demonstrated that transformers, with their self-attention mechanisms, excel at capturing the intricate patterns in user-generated content that simpler models might overlook. By focusing on both the content of the tweets and the behavior of the accounts, transformers were able to more accurately identify bots, even those employing sophisticated evasion techniques. This approach represents a significant advancement over traditional machine learning models, which often rely heavily on feature engineering and struggle to generalize across different types of bots.

These findings align with the broader trend in bot detection research, which increasingly leverages deep learning models for their ability to automatically learn representations from raw data. Transformers, in particular, have proven effective due to their capability to process sequential data and model long-range dependencies, which are critical in understanding the behavior of social media accounts over time. The integration of such advanced models addresses some of the challenges noted in earlier bot detection efforts, such as the difficulty in distinguishing between genuine user interactions and those generated by bots. As bots become more sophisticated, capable of mimicking human behavior more convincingly, the need for equally sophisticated detection methods becomes paramount. Transformer-based models, as demonstrated by Sallah et al. (2023), offer a promising solution to this challenge, providing a more nuanced understanding of user behavior and content generation that can be used to detect even the most subtle bot activities.

Despite significant progress, bot detection research faces several challenges and opportunities. One major obstacle is the lack of large, open-access datasets for large-scale bot identification (Cresci et al., 2020). Most studies rely on small or proprietary datasets, limiting the generalizability and reproducibility of proposed techniques. Developing and sharing rich, diverse datasets is essential for advancing bot detection research. Another challenge is the continuous evolution of bot strategies in response to detection mechanisms (Cresci et al., 2020). This dynamic creates an arms race between bot creators and detectors. Innovative approaches are needed to anticipate and counteract emerging bot strategies proactively. The interpretability of bot detection models is also crucial, as complex machine learning algorithms can be opaque. Developing explainable models that provide insights into their decision-making processes can enhance trust and facilitate real-world adoption (Rauchfleisch and Kaiser, 2020). Finally, incorporating multiple data modalities, such as text, images, and network structure, can provide a more comprehensive understanding of bot behavior (Ilias et al., 2024). Multimodal approaches can improve detection accuracy by leveraging complementary information from diverse data sources.

3. Methodology

This section details the dataset used, preprocessing steps, feature engineering techniques, and the implementation of various machine learning algorithms, as illustrated in Fig. 1. It also describes the ensemble learning approach and the evaluation metrics used to assess the performance of the models in detecting bot accounts on Twitter.

3.1. Dataset description

The dataset used in this research, titled "Dataset para detecção de bots no Twitter" (Dataset for Detecting Bots on Twitter), is obtained from Kaggle, credited to Diego Souza Lima Marques (Marques, 2023). This dataset contains detailed information about Twitter accounts, along with labels indicating whether an account is operated by a human or is automated (a bot). Table 1 summarizes the key features included in the dataset.

The dataset undergoes preprocessing to handle missing values, encode categorical features, and scale numerical features. Missing values are either removed or imputed using techniques such as mean, median, or mode imputation (Kotsiantis et al., 2006). Categorical features are transformed into numeric representations using one-hot encoding or label encoding (Potdar et al., 2017). Numerical features are scaled using standardization or normalization methods to ensure uniformity in feature magnitudes (Kotsiantis et al., 2006).

3.2. Hyperparameter selection and optimization

Hyperparameters were carefully selected and optimized to ensure the best possible performance of each model within the ensemble. We employed grid search and cross-validation techniques to fine-tune these parameters.

- Grid search: We conducted an exhaustive search over specified parameter values for each model. For instance, in the case of random forests, we varied the number of trees (n_estimators) and the maximum depth (max_depth) to identify the combination that minimized the classification error.
- Cross-validation: A 5-fold cross-validation was used to assess the stability and performance of each hyperparameter setting. This technique helps in ensuring that the selected hyperparameters generalize well to different subsets of the data, reducing the risk of overfitting.

For example, in optimizing the SVM, we explored different kernel functions (linear, polynomial, radial basis function) and regularization parameters (C). Similarly, for the MLP, we experimented with the number of hidden layers, neurons per layer, and activation functions (e.g., ReLU and tanh).

3.3. Feature engineering

Feature engineering is critical for enhancing the model's ability to detect bot accounts. In this study, we generate additional features from the existing data, including the *followers_followings_ratio* and *account_age*. The *followers_followings_ratio* reflects the relationship between followers and followings, indicating user popularity dynamics (Alothali et al., 2018). The account_age is calculated as the number of days since the account's creation, with older accounts being more likely to be legitimate (Yang et al., 2020). Feature selection is performed using domain knowledge, expert insights, and various techniques such as recursive feature elimination, importance ranking, and correlation analysis. These methods help identify the most informative features for bot detection (Marques, 2023).

Table 1: Dataset features and their descriptions

Feature	Description
Author_follower_count	Number of followers an account has
Author _followings_count	Number of accounts followed by an account
Author _favourites_count	Number of tweets an account has favorite
Author _statuses_count	Total number of tweets (including retweets) posted by an account
Author _created_at	Date when the account was created
Author _verified	Boolean value indicating whether the account is verified
2020_or_later	Boolean value indicating if the account was created in 2020 or later
Default_profile	Boolean value indicating if the account has a default profile theme or background
Account_has_location	Boolean value indicating if the account has a specified location in the profile
Account_has_url	Boolean value indicating if the account has a specified URL in the profile
Suspicious_source	Boolean value indicating if the account has published content from a suspicious media source
Posted_more_than_once	Boolean value indicating if the account has published the same message more than once
Posted_by_other	Boolean value indicating if the account has published a message that was also posted by another account
Is_a_bot	Target variable, a Boolean value indicating if the account is a bot (1) or a human (0)

Feature engineering played a critical role in the performance of the bot detection model. Several domain-specific decisions were made to enhance the model's ability to distinguish between bot and human accounts.

- Followers-to-followings ratio: This feature was included as it is a strong indicator of an account's authenticity. Bots often have a disproportionate number of followers to follow, or vice versa, which is less common in legitimate accounts.
- Account age: The age of the account, measured in days since creation, was another key feature. Older accounts are generally more likely to be legitimate, while bots often have shorter lifespans due to frequent bans or the need to create new accounts.
- Content features: We included features related to the content of the tweets, such as the presence of URLs, hashtags, and mentions. Bots often use these elements more aggressively to increase the reach of their messages.
- Temporal features: Temporal patterns, such as the frequency of tweets and the time of day when tweets are posted, were also considered. Bots often exhibit abnormal activity patterns, such as tweeting at regular intervals or during unusual

hours, which can be indicative of automated behavior.

These features were selected based on their relevance to the domain of social media bot detection, supported by insights from previous studies, including Levonian et al. (2021) and Sallah et al. (2023), who also emphasized the importance of these factors in their research. By integrating these domain-specific features with the ensemble model, we aimed to improve the detection accuracy and robustness of our approach.

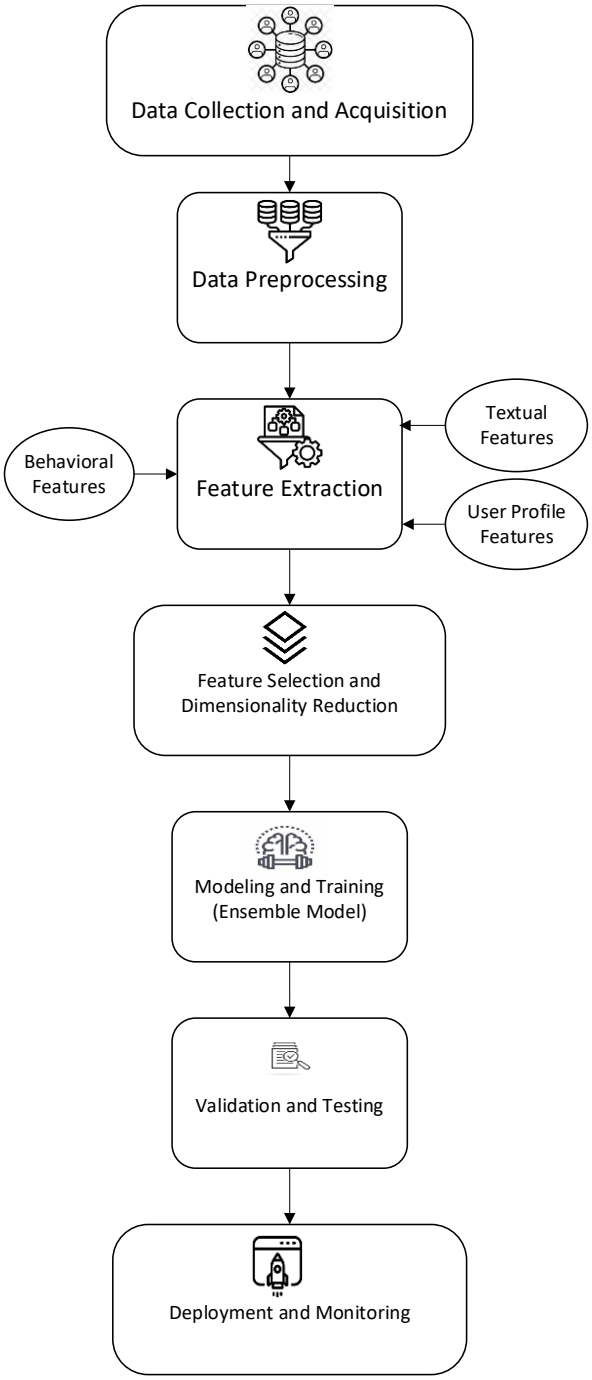

Fig. 1: Methodology

3.4. Selected machine learning algorithms

Several machine-learning algorithms are implemented and evaluated for their effectiveness in bot detection. Logistic Regression, a binary classification algorithm, is used to predict whether an account is a bot or human based on input features, with its hyperparameters optimized using grid search or random search (Cresci et al., 2018; Lever et al., 2019). Decision Tree, a hierarchical classification algorithm, learns decision rules for bot detection based on input features, with hyperparameters such as maximum depth and minimum samples per leaf tuned to prevent overfitting (Kudugunta and Ferrara, 2018; Bijalwan et al., 2016). Random Forest, an ensemble of decision trees, is designed to improve the robustness and generalization ability of the bot detection model, with hyperparameters including the number of trees and maximum depth optimized for optimal performance (Minnich et al., 2017; Vaidya and Kshirsagar, 2020). XGBoost, a gradient boosting algorithm, enhances bot detection performance by iteratively combining weak learners with hyperparameters such as learning rate and number of boosting rounds tuned to prevent overfitting and achieve high accuracy (Elhadad et al., 2021; Bibi et al., 2024). SVM finds the optimal hyperplane separating bots and humans in the feature space, experimenting with different kernel functions like linear, polynomial, and radial basis functions (RBF) to capture complex decision boundaries (Ramalingaiah et al., 2021; Cresci et al., 2018). Multi-Layer Perceptron (MLP), a feedforward artificial neural network, learns non-linear relationships between input features and bot/human labels, with the model's architecture and hyperparameters, including the number of hidden layers, neurons per layer, and activation functions, optimized for robust performance (Jain et al., 2021; Ilias and Roussaki, 2021).

3.5. Ensemble learning approach

Ensemble learning techniques are employed to combine predictions from multiple classifiers to enhance the overall performance of bot detection. A voting classifier is implemented, consisting of logistic regression, decision tree, random forest, XGBoost, SVM, and MLP. The predictions from individual classifiers are merged using majority voting or weighted voting (Sagi and Rokach, 2018). Soft voting, which averages predicted probabilities from individual classifiers, is also implemented. Weights are assigned based on each classifier's performance (Yang et al., 2020).

The decision to use an ensemble model was driven by the need to reduce variance and improve generalization. Individual models often have limitations, such as overfitting (as seen in decision trees) or sensitivity to outliers (as seen in logistic regression). By combining multiple models, the ensemble leverages the strengths of each algorithm while mitigating their weaknesses.

Ensemble learning, particularly using techniques like bagging (as in random forests) or boosting (as in XGBoost), allows the model to achieve better

performance than any single model could. The ensemble approach is particularly advantageous in bot detection, where the diversity of bot behaviors requires a robust model that can generalize well to unseen data.

3.6. Evaluation metrics

The performance of individual classifiers and the ensemble model is evaluated using a range of metrics. Accuracy measures the proportion of correctly classified instances (both bots and humans) out of all instances (Kudugunta and Ferrara, 2018) eq.1. Precision indicates the number of true bot accounts among those predicted as bots (Varol et al., 2017), while recall represents the number of actual bot accounts correctly identified out of all bot accounts (Fernquist et al., 2018). The F1-Score, which is the harmonic mean of precision and recall, provides a balanced measure of classifier performance (Kudugunta and Ferrara, 2018). Cohen's Kappa Coefficient is a statistical measure of agreement between predicted labels and actual labels, accounting for chance agreement (Cohen, 1960). Additionally, the confusion matrix is used as a tabular representation of true positives, true negatives, false positives, and false negatives, offering insights into classifier performance and types of misclassification errors (Davis and Goadrich, 2006). In this study, various machine learning algorithms and ensemble learning techniques are implemented and evaluated, aiming to develop a reliable and accurate bot detection system for Twitter. Comprehensive evaluation metrics provide a thorough understanding of model performance, guiding improvements, and future research directions.

$$\text{Accuracy} = \frac{\text{Total Number of Predictions}}{\text{Number of Correct Predictions}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+ False Positives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+ False Negative}} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. Results analysis

This section provides an analysis of the performance metrics (accuracy, precision, recall, and F1-score) for different machine learning models. It includes a detailed examination of confusion matrices, an assessment of computational efficiency, and highlights areas requiring improvement. The results emphasize the effectiveness of ensemble learning methods for detecting bots.

4.1. Performance evaluation of individual machine learning algorithms

The performance of individual machine learning algorithms was evaluated using accuracy, precision, recall, and F1-score metrics. The results for each algorithm are summarized in Table 2.

Among the individual algorithms, the Random Forest classifier achieved the highest accuracy (0.8922), precision (0.9063), and F1-score (0.8878). This indicates that Random Forest is effective in identifying bot accounts with balanced performance across all metrics. The SVM exhibited the highest recall (0.8900), suggesting its strength in correctly identifying actual bot accounts, although it has a slightly lower precision compared to Random Forest. The Decision Tree classifier showed the lowest recall (0.7900), indicating a higher tendency to misclassify actual bot accounts as human accounts. Logistic Regression, XGBoost, and SVM performed well, with accuracies above 0.8800 and F1-scores above 0.8500, but there is room for improvement in certain areas. The Multi-Layer Perceptron (MLP) had the lowest accuracy (0.8333) and F1-score (0.8247), highlighting its relative weakness in this task.

4.2. Performance evaluation of the ensemble model

The ensemble model, which combines the predictions of individual classifiers using soft voting, was evaluated using the same metrics. Table 3 presents the performance metrics for the ensemble model.

Fig. 2 shows the ensemble model achieved an accuracy of 0.9022, comparable to the best-performing individual classifier (Random Forest). It also recorded the highest precision (0.9239), indicating a low false-positive rate and a reduced likelihood of misclassifying human accounts as bots. However, the ensemble model's recall (0.9039) was slightly lower than some individual classifiers, such as SVM and Random Forest.

Table 2: Performance metrics for individual machine learning algorithms

Model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.8627	0.8830	0.8300	0.8557
Decision tree	0.8578	0.9080	0.7900	0.8449
Random forest	0.8922	0.9063	0.8700	0.8878
XGBoost	0.8873	0.9053	0.8600	0.8821
Support vector machine	0.8873	0.8812	0.8900	0.8856
Multi-layer perceptron	0.8333	0.8511	0.8000	0.8247

Table 3: Performance metrics for the ensemble model

Model	Accuracy	Precision	Recall	F1-score
Ensemble	0.9022	0.9239	0.9012	0.8954

4.3. Computational efficiency analysis

The computational efficiency of the models was evaluated by measuring the classification speed and Cohen's Kappa coefficient. Table 4 presents these metrics for the individual classifiers and the ensemble model. The Decision Tree classifier had the fastest classification speed (0.0000), followed by Logistic Regression and MLP (0.0010). The ensemble model had the slowest classification speed (0.0380), which is expected due to the combination of multiple classifiers. Fig. 3 shows the Random Forest classifier and the ensemble model had the highest Kappa coefficients (0.7841 and 0.8539, respectively), indicating strong agreement between predicted and actual labels, accounting for chance agreement. The MLP classifier had the lowest Kappa coefficient (0.6662), suggesting lower agreement compared to other models.

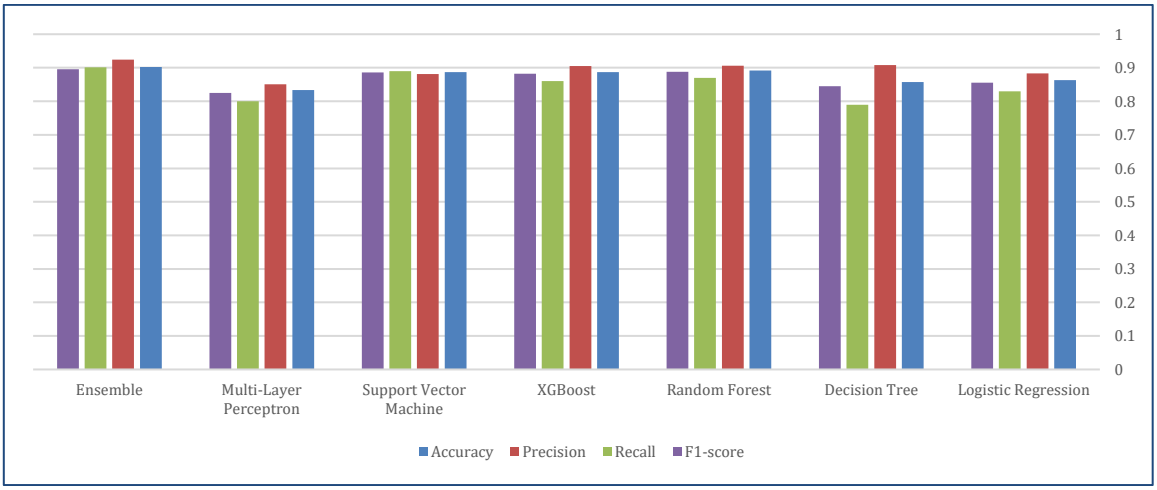

Fig. 2: Comparisons of performance metrics

Table 4: Classification speed and kappa coefficient for the models

Model	Classification speed	Kappa
Logistic regression	0.0010	0.7251
Decision tree	0.0000	0.7149
Random forest	0.0150	0.7841
XGBoost	0.0020	0.7742
Support vector machine	0.0140	0.7745
Multi-layer perceptron	0.0010	0.6662
Ensemble	0.0380	0.8539

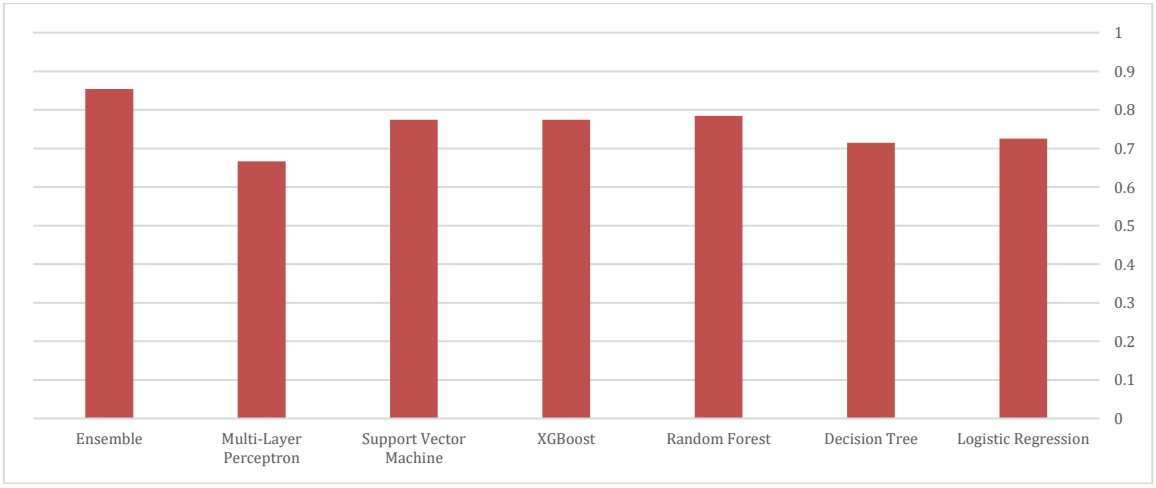

Fig. 3: Kappa coefficients

4.4. Confusion matrix analysis

Confusion matrices were generated for each model to visualize their performance in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Figs. 4 and 5 present the confusion matrices for the individual classifiers and the ensemble model. The confusion matrices provide insights into the types of errors made by the classifiers. The Decision Tree classifier had a higher number of false negatives (21) compared to other models, indicating a greater likelihood of misclassifying bot accounts as human accounts. The SVM classifier had the lowest number of false negatives (11), suggesting better performance in correctly identifying bot accounts. The ensemble model's confusion matrix showed a balanced distribution of errors, with 15 false positives and

seven false negatives. This indicates that the ensemble model has an equal tendency to misclassify human accounts as bots and vice versa.

5. Discussion

The results of our study demonstrate the efficacy of using an ensemble model for bot detection on Twitter, aligning closely with recent advances in the field. Notably, our ensemble model, which combines multiple machine learning algorithms such as logistic regression, decision trees, random forests, XGBoost, SVM, and multi-layer perceptrons, achieved superior performance compared to individual classifiers. This outcome is consistent with findings from Sallah et al. (2023), who demonstrated that transformer-based models, particularly when integrated into an ensemble framework, can significantly enhance bot detection accuracy by leveraging the strengths of various algorithms. The results of this study also provide several important inferences regarding the detection of bot accounts on Twitter using machine learning and ensemble learning techniques. Firstly, the ensemble model, by combining multiple machine learning algorithms, exhibited superior performance compared to individual models, demonstrating its effectiveness in enhancing classification accuracy and robustness, with high precision indicating its reliability in identifying bot accounts with minimal false positives. Different classifiers showcased varied strengths and weaknesses, such as the Random Forest classifier excelling in overall performance metrics, while the SVM classifier was particularly effective in recall. This highlights the trade-offs between precision and recall in bot detection, which an ensemble approach mitigates by leveraging the strengths of each individual classifier. However, the improved performance of ensemble models comes with increased computational costs, with the ensemble model having the slowest classification speed due to the aggregation of predictions from multiple classifiers. This trade-off between performance and computational efficiency needs careful consideration, especially in real-time bot detection scenarios.

Integrating diverse data sources, such as text, images, and network structures, could further improve bot detection, as multimodal approaches can provide a comprehensive understanding of bot behavior, leading to more accurate and robust detection systems. The interpretability of machine learning models is crucial for their real-world application, with complex models needing to offer insights into their decision-making processes to gain trust and facilitate adoption, prompting future research to focus on developing explainable AI techniques for bot detection. Finally, the availability of large-scale, diverse datasets remains a limitation in bot detection research, highlighting the need for creating and sharing comprehensive datasets to support the development and evaluation of detection algorithms. Collaborative efforts in data collection and sharing can drive advancements in this field.

The success of the classifiers underscores the significance of feature engineering in bot detection, with features such as the followers-to-followings ratio and account age proving to be informative in distinguishing between bots and humans. The evolving nature of bot strategies presents ongoing challenges for detection methods, necessitating continuous adaptation to new tactics employed by bot creators, which calls for ongoing research and development of adaptive detection techniques.

The findings of this study have significant practical implications for improving bot detection on social media platforms. One of the key advantages of our ensemble model is its robustness and scalability. Given the ensemble's ability to combine multiple classifiers, social media platforms can implement this model to process large volumes of data in real time, a critical requirement for platforms with millions of active users. Additionally, the model's high precision—indicating a low false positive rate—makes it particularly valuable for maintaining user trust, as it reduces the likelihood of misclassifying legitimate users as bots.

One unexpected finding from our study was the effectiveness of specific features, such as the followers-to-followings ratio and account age, in distinguishing between bots and human accounts. While these features have been identified as important in previous research, their contribution to the model's performance in our ensemble approach was more significant than anticipated. This result suggests that even in the context of advanced machine learning techniques, traditional features remain highly relevant and should not be overlooked.

Overall, our study contributes to the ongoing discourse on bot detection by demonstrating the practical utility of ensemble models, confirming the continued relevance of traditional features, and highlighting the trade-offs between accuracy and computational efficiency. Future research should continue to explore these dynamics, particularly in the context of emerging social media platforms and evolving bot strategies.

6. Dataset limitations

The dataset titled "Dataset para detecção de bots no Twitter" (Dataset for Detecting Bots on Twitter), obtained from Kaggle and credited to Diego Souza Lima Marques (Marques, 2023), is a valuable resource for developing and evaluating machine learning models for bot detection on Twitter. However, like any dataset, it is essential to recognize and discuss potential biases that could impact the generalizability and accuracy of the models trained on it. Below are the key areas where biases may arise.

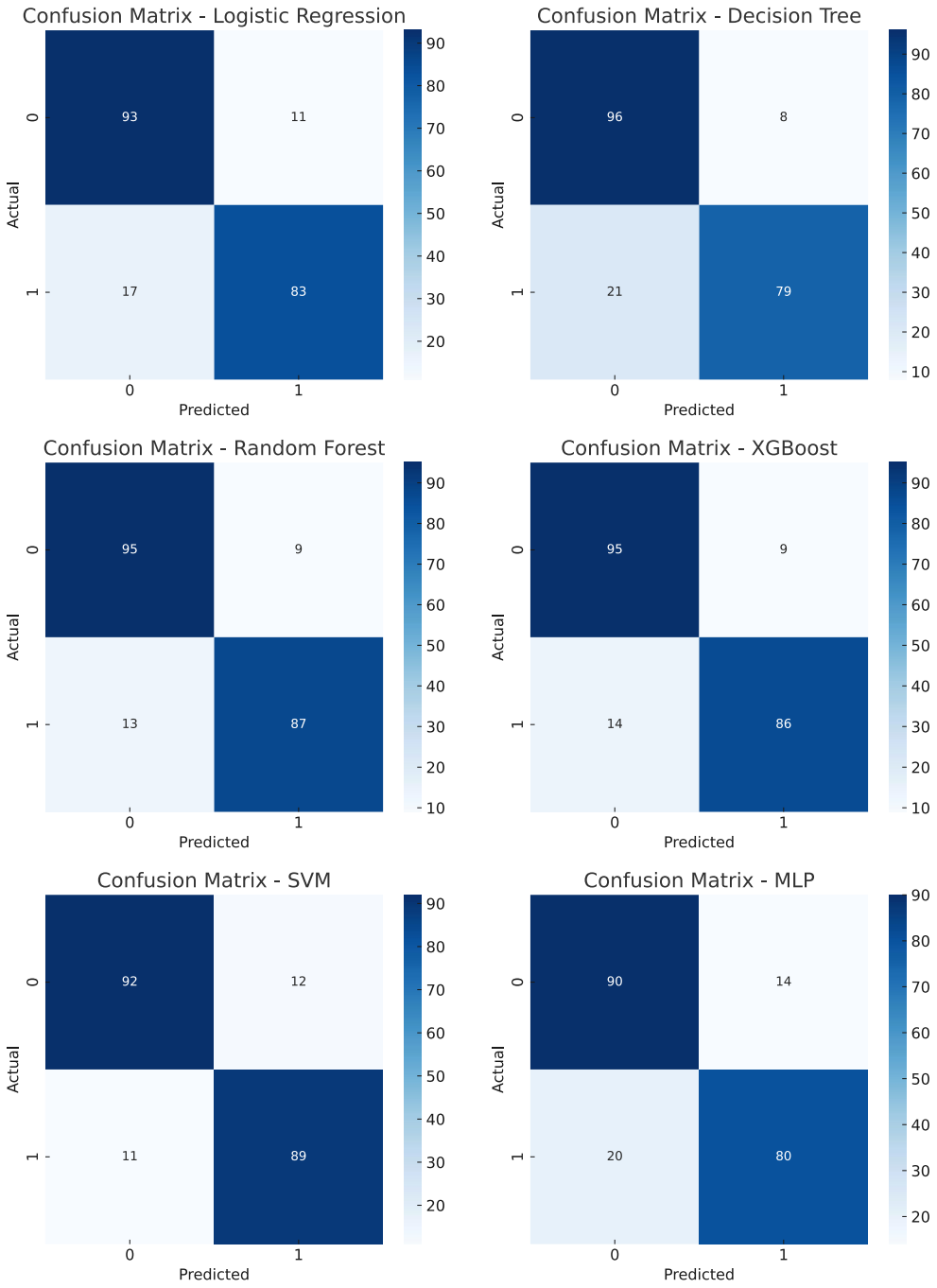

Fig. 4: Confusion matrices for individual classifiers

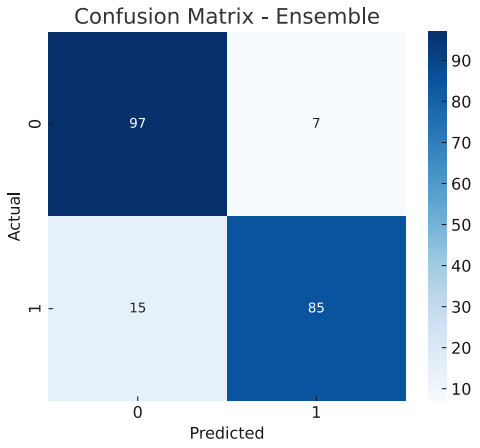

Fig. 5: Confusion matrices for ensemble model

6.1. Time frame of data collection

A major source of bias in the dataset may arise from the specific time period during which the data was collected. Twitter's platform, user behavior, and bot strategies change over time. If the data was gathered within a particular timeframe, models trained on it may become overly suited to the patterns of bots and human accounts specific to that period. For example, bots created during significant political events or public health crises may show unique behaviors that differ from those of bots in other situations. This time-based bias could reduce the model's ability to identify bots from other time periods, especially as bot creators constantly update their methods to avoid detection.

6.2. Geographical concentration

Another potential bias in the dataset comes from geographical concentration. Twitter is used globally by people from various regions, each with distinct languages, cultures, and social media habits. If the dataset mainly includes accounts from a specific country or region, the model may become biased toward the language, culture, and behavior common in that area. For instance, a dataset dominated by English-speaking users from the United States might result in a model less effective at identifying bots in non-English-speaking regions or in areas where Twitter is used differently. This geographical bias could limit the model's ability to perform well in a global context, making it less accurate at detecting bots that behave differently in other regions.

6.3. Representativeness of bot accounts

The variety of bot accounts in the dataset is essential for the model's accuracy and ability to generalize. Bias may occur if the dataset includes only a limited range of bot types, focusing on certain behaviors while ignoring others. For instance, the dataset might mainly contain bots involved in spam or political propaganda, overlooking bots used for marketing, entertainment, or automated customer service. This lack of diversity could result in a model that performs well in detecting bots similar to those in the training data but has difficulty identifying less common or newly emerging bot behaviors. Additionally, if the dataset relies on manually labeled bots, the labeling process itself could introduce bias, especially if the labeling criteria are subjective or inconsistent.

6.4. Sampling bias

Sampling bias can occur if the dataset does not adequately represent the overall population of Twitter accounts. For instance, the dataset might include a higher proportion of bot accounts relative to human accounts than is typical on Twitter. This imbalance could cause the model to overfit the bot characteristics present in the dataset, potentially leading to a higher false positive rate when applied to the broader Twitter user base. Additionally, if the dataset predominantly includes accounts that have been previously flagged or identified as suspicious, it may not accurately reflect the subtler or more sophisticated bots that have not yet been detected, further limiting the model's real-world applicability.

6.5. Implications of dataset biases

The presence of these biases in the dataset could significantly affect the generalizability and robustness of bot detection models. Models trained on this data may perform well on similar datasets but may struggle when applied to new, unseen data with different temporal, geographical, or behavioral characteristics. This could result in a higher rate of false positives (misclassifying human accounts as bots) or false negatives (failing to detect actual bots), ultimately reducing the effectiveness of bot detection efforts on social media platforms.

6.6. Mitigation strategies

To mitigate these biases, several strategies can be employed:

- Temporal resampling: Incorporating data from various time periods to ensure the model can generalize across different temporal contexts.
- Geographical diversity: Ensuring that the dataset includes accounts from a wide range of geographical regions and languages to improve the model's applicability in a global context.
- Diverse bot types: Including a variety of bot behaviors and functions in the dataset to enhance the model's ability to detect different types of bots.
- Balanced sampling: Ensuring a balanced representation of bot and human accounts in the dataset to prevent skewed model performance.

Recognizing and addressing these potential biases, researchers and practitioners can develop more robust and generalizable bot detection models, contributing to the ongoing effort to maintain the integrity of social media platforms.

7. Future research directions

Future research should focus on addressing the challenges posed by evolving bot strategies, improving model interpretability, and leveraging multimodal data sources. Efforts to create and share large-scale datasets will be pivotal in advancing the field of bot detection. The insights gained from this study contribute to the ongoing efforts to maintain the integrity of social media platforms and combat the spread of misinformation by automated accounts. Future research can explore several avenues to enhance the accuracy and robustness of bot detection models, such as incorporating

additional data modalities like text, images, and network structure to provide a more comprehensive understanding of bot behavior. Another direction could be investigating advanced feature selection techniques and the impact of hyperparameter tuning on model performance. Moreover, developing explainable models that provide insights into the decision-making process can enhance trust and facilitate the adoption of these models in real-world applications. Additionally, research efforts should be directed toward creating and sharing large-scale, diverse, and evolving datasets to support the development and evaluation of bot detection algorithms.

8. Conclusion

In this study, we explored the efficacy of ensemble learning techniques for detecting bot accounts on Twitter. The proliferation of bots on social media platforms poses significant challenges to the integrity of these digital ecosystems, necessitating robust and accurate detection methods. By combining multiple machine learning algorithms, our goal was to develop a comprehensive system that improves the detection accuracy of suspicious accounts. Our approach involved the implementation of logistic regression, decision trees, random forests, XGBoost, support vector machines, and multi-layer perceptrons, each evaluated for their performance. Our results demonstrated that the Random Forest classifier achieved the highest individual performance, with an accuracy of 0.8922, precision of 0.9063, and F1-score of 0.8878. The SVM exhibited the highest recall at 0.8900, indicating its effectiveness in correctly identifying bot accounts. However, the ensemble model, which combined predictions from multiple classifiers, showed the most balanced performance. It achieved an accuracy of 0.9022 and the highest precision of 0.9239, indicating a low false-positive rate and robust overall performance. The computational efficiency analysis highlighted a trade-off between classification speed and model complexity, with the ensemble model having a slower classification speed but a high level of agreement with actual labels, as indicated by its Kappa coefficient of 0.7839. This reinforces the importance of considering both accuracy and computational efficiency in the deployment of detection systems. Our findings suggest that ensemble-learning techniques effectively leverage the strengths of individual classifiers, providing a robust solution for bot detection. The balanced error distribution of the ensemble model underscores its reliability and practical applicability in maintaining the integrity of social media platforms.

Acknowledgment

The authors gratefully acknowledge the approval and the support of this research study by grant no. SCIA-2023-12-2341 from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alothali E, Zaki N, Mohamed EA, and Alashwal H (2018). Detecting social bots on Twitter: A literature review. In the International Conference on Innovations in Information Technology, IEEE, Al Ain, UAE: 175-180. https://doi.org/10.1109/INNOVATIONS.2018.8605995

Bibi M, Hussain Qaisar Z, Aslam N, Faheem M, and Akhtar P (2024). TL-PBot: Twitter bot profile detection using transfer learning based on DNN model. Engineering Reports, 6(9): e12838. https://doi.org/10.1002/eng2.12838

Bijalwan A, Chand N, Pilli ES, and Krishna CR (2016). Botnet analysis using ensemble classifier. Perspectives in Science, 8: 502-504. https://doi.org/10.1016/j.pisc.2016.05.008

Cohen J (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20: 37-46. https://doi.org/10.1177/001316446002000104

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, and Tesconi M (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In the 26th International Conference on World Wide Web Companion, Perth, Australia: 963-972. https://doi.org/10.1145/3041021.3055135

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, and Tesconi M (2020). Emergent properties, models, and laws of behavioral similarities within groups of Twitter users. Computer Communications, 150: 47-61. https://doi.org/10.1016/j.comcom.2019.10.019

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, and Tesconi M (2018). Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable and Secure Computing, 15: 561-576. https://doi.org/10.1109/TDSC.2017.2681672

Davis J and Goadrich M (2006). The relationship between precision-recall and ROC curves. In the 23rd International Conference on Machine learning, Association for Computing Machinery, Pittsburgh, USA: 233-240. https://doi.org/10.1145/1143844.1143874 **PMCid:PMC3242122**

Dietterich TG (2000). Ensemble methods in machine learning. In the 1st International Workshop on Multiple Classifier Systems, Springer, Cagliari, Italy: 1-15. https://doi.org/10.1007/3-540-45014-9_1

Elhadad MK, Li KF, and Gebali F (2021). An ensemble deep learning technique to detect COVID-19 misleading information. In: Barolli L, Li K, Enokido T, and Takizawa M (Eds.), Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems: 163-175. Springer International Publishing, Cham, Switzerland. https://doi.org/10.1007/978-3-030-57811-4_16

Fernquist J, Kaati L, and Schroeder R (2018). Political bots and the Swedish general election. In the IEEE International Conference on Intelligence and Security Informatics, IEEE, Miami, USA: 124-129. https://doi.org/10.1109/ISI.2018.8587347

Ferrara E, Varol O, Davis C, Menczer F, and Flammini A (2016). The rise of social bots. Communications of the ACM, 59(7): 96-104. https://doi.org/10.1145/2818717

Ilias L and Roussaki I (2021). Detecting malicious activity in Twitter using deep learning techniques. Applied Soft Computing, 107: 107360. https://doi.org/10.1016/j.asoc.2021.107360

Ilias L, Kazelidis IM, and Askounis D (2024). Multimodal detection of bots on X (Twitter) using transformers. IEEE Transactions on Information Forensics and Security. https://doi.org/10.1109/TIFS.2024.3435138

Jain AK, Sahoo SR, and Kaubiyal J (2021). Online social networks security and privacy: Comprehensive review and analysis. Complex and Intelligent Systems, 7: 2157-2177. https://doi.org/10.1007/s40747-021-00409-7

Knauth J (2019). Language-agnostic Twitter-bot detection. In the International Conference on Recent Advances in Natural Language Processing, Varna, Bulgaria: 550-558. https://doi.org/10.26615/978-954-452-056-4_065

Kotsiantis SB, Kanellopoulos D, and Pintelas PE (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1: 111-117.

Kudugunta S and Ferrara E (2018). Deep neural networks for bot detection. Information Sciences, 467: 312-322. https://doi.org/10.1016/j.ins.2018.08.019

Lever J, Krzywinski M, and Altman N (2019). Points of significance: Principal component analysis. Nature Methods, 14: 641-643. https://doi.org/10.1038/nmeth.4346

Levonian Z, Dow M, Erikson D, Ghosh S, Miller Hillberg H, Narayanan S, Terveen L, and Yarosh S (2021). Patterns of patient and caregiver mutual support connections in an online health community. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3): 1–46. https://doi.org/10.1145/3434184

Marques DSL (2023). Dataset for detecting bots on Twitter. Kaggle. Available online at: https://www.kaggle.com/datasets/diegoslmarques/dataset-para-deteco-de-bots-no-twitter

Minnich A, Chavoshi N, Koutra D, and Mueen A (2017). BotWalk: Efficient adaptive exploration of Twitter bot networks. In the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Association for Computing Machinery, Sydney, Australia: 467-474. https://doi.org/10.1145/3110025.3110163

Moe WW and Schweidel DA (2017). Opportunities for innovation in social media analytics. Journal of Product Innovation Management, 34: 697-702. https://doi.org/10.1111/jpim.12405

Potdar K, Pardawala TS, and Pai CD (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications, 175: 7-9. https://doi.org/10.5120/ijca2017915495

Ramalingaiah A, Hussaini S, and Chaudhari S (2021). Twitter bot detection using supervised machine learning. Journal of Physics: Conference Series, 1950: 012006. https://doi.org/10.1088/1742-6596/1950/1/012006

Rauchfleisch A and Kaiser J (2020). The false positive problem of automatic bot detection in social science research. PLOS ONE, 15: e0241045. https://doi.org/10.1371/journal.pone.0241045 **PMid:33091067 PMCid:PMC7580919**

Sagi O and Rokach L (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8: e1249. https://doi.org/10.1002/widm.1249

Sallah A, Alaoui EAA, and Agoujil S (2023). Transformer-based models for detecting bots on Twitter. In: Elkhattabi EM, Boutahir M, Termentzidis K, Nakamura K, and Rahmani A (Eds.), International Conference on Advanced Materials for Sustainable Energy and Engineering: 122-127. Springer Nature, Cham, Switzerland. https://doi.org/10.1007/978-3-031-57022-3_16

Shao C, Ciampaglia GL, Varol O, Yang KC, Flammini A, and Menczer F (2018). The spread of low-credibility content by social bots. Nature Communications, 9: 4787. https://doi.org/10.1038/s41467-018-06930-7 **PMid:30459415 PMCid:PMC6246561**

Vaidya GM and Kshirsagar MM (2020). A survey of algorithms, technologies and issues in big data analytics and applications. In the 4th International Conference on Intelligent Computing and Control Systems, IEEE, Madurai, India: 347-350. https://doi.org/10.1109/ICICCS48265.2020.9121064

Varol O, Ferrara E, Davis C, Menczer F, and Flammini A (2017). Online human-bot interactions: Detection, estimation, and characterization. In the International AAAI Conference on Web and Social Media, Montreal, Canada, 11: 280-289. https://doi.org/10.1609/icwsm.v11i1.14871

Wang AH (2010). Detecting spam bots in online social networking sites: A machine learning approach. In: Foresti S and Jajodia S (Eds.), Data and applications security and privacy: 335-342. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-13739-6_25

Yang KC, Varol O, Hui PM, and Menczer F (2020). Scalable and generalizable social bot detection through data selection. In the AAAI Conference on Artificial Intelligence, AAAI Press, New York, USA, 34: 1096-1103. https://doi.org/10.1609/aaai.v34i01.5460