Detection of Malicious Social Bots using Generative Adversarial Networks with Recurrent Neural Networks in Twitter Network

Greeshma Lingam, Bhavya Yasaswini, Ratna Priya and Niharika
greeshma243@gmail.com

***Abstract*—Malicious social bots are the propagators who spread malicious information in online social networks (OSNs). Thus, the malicious social bots affect OSN environment and information security. In OSNs, the number of malicious social bots is much less than legitimate users. Therefore, the detection of malicious social bot with better accuracy is one of the important research topic. Most of the existing social bot detection approaches rely on supervised learning models. However, most of the conventional supervised learning models suffer from highly unbalanced dataset with many legitimate users belonging to one class and only a few malicious social bots belonging to another class. In order to solve the problem of unbalanced distribution of malicious social bot, a generative adversarial network with recurrent neural network (GAN-RNN) model has been proposed. Experimentation have been performed on *Twitter* dataset in order to verify the efficacy of the proposed GAN-RNN model.**

Keywords*—*Malicious Social Bots, Unbalanced data, Generative Adversarial Networks, Recurrent Neural Network, Twitter network.

I. INTRODUCTION

Online social networks (OSNs) have influenced users to share information related to social activities like news, links, opinion and promote product and services. Online social networking sites contain huge amount of data (such as data from online reviews, online ratings and discussions forums) which are generated by users (from various communities) [1]. The data can be accessed seamlessly due to proliferation of online social network technologies. This in turn provides an additional space (or comfort) for an attacker to steal user's personal information and to perform malicious activities (such as generating fake identities, manipulating online ratings, spreading social spam content and performing phishing attacks) in online social networks [2]. Therefore, in recent years, malicious social bots are the major threats in OSNs.

Malicious social botnet is a group of malicious social bots, where each malicious social bot represent is a software program which automatically creates multiple fake accounts, frequently interacts with legitimate user accounts [3]. Further, malicious social botnet tries to disseminate social spam content (or fake information) with malicious intention and post malicious (or phishing) links in the tweet. For example, malicious social bots can be created using Twitter API [4]. Moreover, such type of malicious social bots can affect OSN environment with several vulnerabilities. In many OSNs, malicious social bots are much less in number when compared to the legitimate users. Most of the existing studies on social bot detection rely on supervised learning models [5], [6]. However, most of the conventional supervised learning models often suffer from highly imbalanced dataset with many legitimate users belonging to one class and only a few malicious social bots belonging to another class [7]. Therefore, the detection of malicious social bot with better accuracy is one of the challenging tasks.

In order to solve imbalanced distribution of malicious social bot, a generative adversarial network with recurrent neural network (GAN-RNN) model has been proposed. We then propose two algorithms namely, long short term memory recurrent neural network with GAN (LSTM-GAN) and gated recurrent unit with GAN (GRU-GAN) to identify the malicious social bots in Twitter network. In GAN, the generator and discriminator are modeled using LSTM and GRU. In GAN, the generator generates fake data samples by incorporating random labeled data with noise and this leads to form an augmented balanced training dataset. Further, the discriminator has to distinguish benign users from malicious social bots. This is a challenging task for the discriminator to identify the malicious social bots more accurately. The major contribution of this paper is summarized as follows:

- Analyze the behavior of user by considering user-based features such as number of favorites, number of retweets, tweet source, number of user mentions and links.
- Design a malicious social bot detection model by integrating a generative adversarial network with two recurrent neural network models namely long short term memory (LSTM) and gated recurrent unit (GRU).
- Experimentation have been performed on *Twitter* dataset in order to verify the efficacy of the proposed model.

The remaining part of this paper is organized as follows: Section 2 discusses about related work. In Section 3, a generative adversarial network with recurrent neural network (GAN-RNN) model has been proposed. In Section 4, experimentation have been performed on *Twitter* dataset. Finally, the paper is concluded in Section 5.

II. RELATED WORK

In this section, analysis and identification of bots in social networks have been discussed.

Rout et al. [8] have analyzed the social botnet behavior using URL features in Twitter network. The authors have considered

Learning Automata (LA) model, a reinforcement learning technique to understand the temporal behavior pattern of social bots at different time slots to detect malicious bots among them. The LA algorithm has been integrated with trust model to compute trust value of a user. The authors did not consider noise removal technique for data processing. Wu et al. [9] have proposed an improved conditional generative adversarial network (improved CGAN) by density peak clustering based with gradient penalty to overcome imbalances in the data samples from Twitter. Further, the authors have considered the generation of data-augmentation noise to improve the detection accuracy of social bots. Wang et al. [10] have proposed a bot detection framework based on a combination of a Variational Auto Encoder and an anomaly detection algorithm. Auto Encoders are used to automatically encode and decode sample features. Anomaly detection method is based on k-nearest neighbors which is used to reduce the number of abnormal samples involved in model training.

Zhang et al. [11] have demonstrated the effectiveness and advantages of exploiting a social botnet for spam distribution and digital-influence manipulation on Twitter. They have also proposed two countermeasures to defend against the two reported attacks considering the approach that a bot only retweets the spam tweets which are posted by botmaster and manipulates the influence value of each user. Evaluation of the performance of the digital-influence measurement scheme has been carried out by using the Twitter geo-search API to collect the users in a specific metropolitan area and then crawling the latest tweets of each user to extract the interactions such as retweets, replies, and mentions. The limitations in this approach are network features are not taken into consideration. Kuduguntaet al. [12] have proposed a deep neural network based on contextual long short-term memory (LSTM) architecture that exploits both content and metadata features to detect bots at the tweet level. Moreover, the contextual features are extracted from user metadata and fed as auxiliary input to LSTM deep nets in order to process the tweet text. The authors did not consider the dynamic behavior of bots. Lingam et al. [13] have proposed a social botnet detection algorithm by incorporating a trust model for identifying a trustworthy path in an online social network. Further, the trust value is determined using Bayesian theory and Dempster-Shafer theory. Shi et al. [14] have presented a novel method of detecting malicious social bots by including both features selection based on the transition probability of clickstream sequences and semi-supervised clustering which analyzes transition probability of user behavior based on clickstreams.

III. Generative Adversarial Networks with Recurrent Neural Networks

The proposed Generative Adversarial Networks with Recurrent Neural Networks model is shown in Fig 1 for bot detection. The proposed model consists of three modules namely Data Collection, Data Balancing with GAN and Classification.

A. Data Collection

From an online social network, we need to extract the user profile based features such as statuses count, followers count

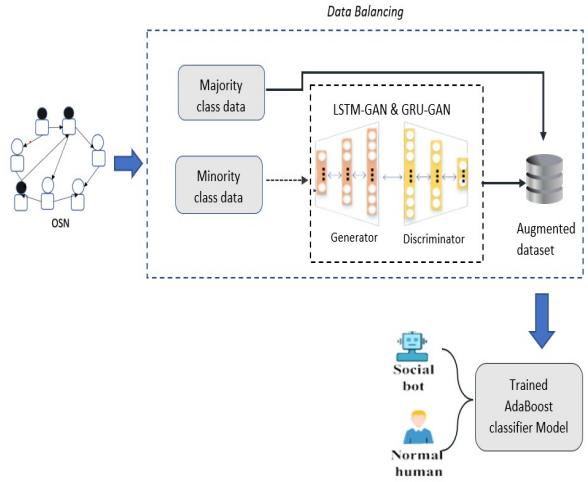


Fig. 1. Proposed GAN with LSTM and GRU Recurrent Neural Networks model

friends count, favourites count and listed count. These features are fed as input to the data balancing model. The data balancing model contains long short term memory recurrent neural network with GAN (LSTM-GAN) and gated recurrent unit with GAN (GRU-GAN) to identify the malicious social bots in Twitter network. In GAN, the generator and discriminator are modeled using LSTM and GRU. In GAN, the generator generates fake data samples by incorporating random labeled data with noise and this leads to form an augmented balanced training dataset. Further, the discriminator distinguishes benign users from malicious social bots for classifying the online social networking accounts.

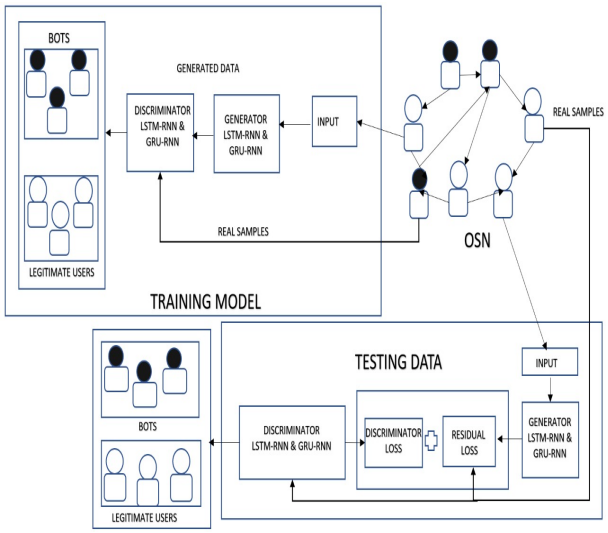


Fig. 2. Workflow of the Proposed model

Algorithm 1 LSTM-GAN and GRU-GAN based Malicious Social Bot Detection Algorithm

Input:
A: Training dataset with Online social networking accounts
Output:
a set of f bots
begin
for (i=1;i¡itr;i++) **do**
Generates data samples from training dataset
$G_{LSTM}(S) = S = \{s_j, j = 1, ...n\}$
$G_{GRU}(S) = S = \{s_j, j = 1, ...n\}$
Compute Discrimination:
$P = \{p_j, j = 1...n\} => D_{LSTM}(P)$
$P' = \{p'_j, j = 1...n\} => D_{GRU}(P')$
$G_{LSTM}(S) => D_{LSTM}(G_{LSTM}(S))$
$G_{GRU}(P) => D_{GRU}(G_{GRU}(P))$
Minimizing Discriminator Loss D_{loss}^{LSTM}
$D_{Loss}^{LSTM} = min\frac{1}{n}\sum_{j=1}^{n}[-logD_{LSTM}p_j - log(1 - D_{LSTM}(G_{LSTM}(s_j))]$
$D_{Loss}^{GRU} = min\frac{1}{n}\sum_{j=1}^{n}[-logD_{GRU}(P_j - log(1 - D_{GRU}(G_{GRU}(s_j))]$
Minimize Generator Loss G_{Loss}^{LSTM} and G_{Loss}^{GRU}
$G_{Loss}^{LSTM} = min\sum_{j=1}^{n} log(-D_{LSTM}(G_{LSTM}(s_j))$
$G_{Loss}^{GRU} = min\sum_{j=1}^{n} log(-D_{GRU}(G_{GRU}(s_j))$
store $D_{Loss}^{LSTM}, D_{Loss}^{GRU}, G_{Loss}^{LSTM}$ and G_{Loss}^{GRU} at each iterator
end for
for each k^{th} iterator **do**
$S^k = min_s Er(P^{test}, G_{LSTM}(S^i))$
$S^{k\prime} = min_s Er(P^{test}, G_{GRU}(S^i))$
end for
Compute residuals:
$r = |p^{test} - G_{LSTM}(S^i)|$
$r' = |p^{test} - G_{GRU}(S^i)|$
Compute discriminator result:
$dr = D_{LSTM}(P^{test})$
$dr' = D_{GRU}(P^{test})$
$R = \alpha r + (1 - \alpha)dr$
$R' = \alpha r' + (1 - \alpha)dr'$
$T = average(R, R')$

B. Data Balancing

In most of the online social networks, the bots are less in number when compared to humans. This may lead to a imbalance problem in an online social network for bot detection [7]. The data can be balanced either by under sampling or oversampling techniques [**?**]. Under sampling is not suitable because of the loss of information in a social-bot-detection scenario which effects the model performance. Data augmentation is applied to generate new fake samples of data from the existing data which is termed as oversampling. GANs discover the regularities in input data in such a way that the model can be used to generate or output new examples that plausibly have been considered from the original dataset. We use two sub-models namely the generator model which is trained to generate new examples, and the discriminator model that tries to classify examples as either real or fake.The generator function generates a batch of fake samples. Further, the fake samples are passed to the discriminator along with the real samples for data classification as real or fake data. The generator and discriminator functions are built with Convolutional neural network layers. Besides neural networks, other structures can be used as discriminative models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). The standard GAN loss function contains two parts such as Generator loss and Discriminator Loss. The discriminator classifies both real and fake data. Further, it penalizes itself for misclassifying a real instance as fake, or a fake instance as real. This is termed as discriminator loss. The generator loss depends on the discriminator classification. LSTMs are a special kind of RNN and it is capable of learning long-term dependencies, remembering information for long periods of time. GRUs are improved version of standard recurrent neural network which are simpler, trains faster than LSTM and have lesser parameter complexity. The proposed long short term memory (LSTM) and gated recurrent unit (GRU) with GAN algorithm has been presented in Algorithm 1. The overall workflow of the proposed model is shown in figure 2.

C. Classification

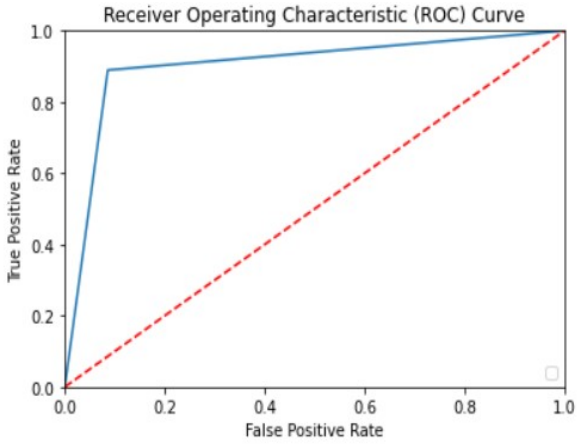
The balanced data from GAN model is passed to Random Forest with adaboost classifier for identification of bots. Boosting is a general ensemble method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. The models are added until the training set is predicted more accurately. AdaBoost is used to boost the performance of decision trees on binary classification problem. It helps to combine multiple weak classifiers into a single strong classifier.

TABLE I. PERFORMANCE EVALUATION TABLE

Model	Recall	Precision	Accuracy
Random Forest with Adaboost	88.60%	93.15%	89.51%
CNN based GAN	88.94%	91.71%	89.51%
LSTM based GAN	89.91%	96.96%	95.32%
GRU based GAN	88.29%	96.95%	95.62%

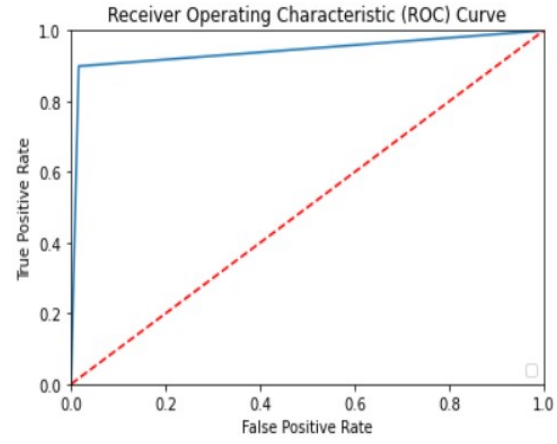
TABLE II. SUMMARY OF DATASETS

Dataset	No. of records
genuine accounts	3474
social spambots#1	991
social spambots#2	3457
Social spambots#3	464
traditional spambots#1	1000
traditional spambots#2	100
traditional spambots#3	403
Traditional spambots#4	1128



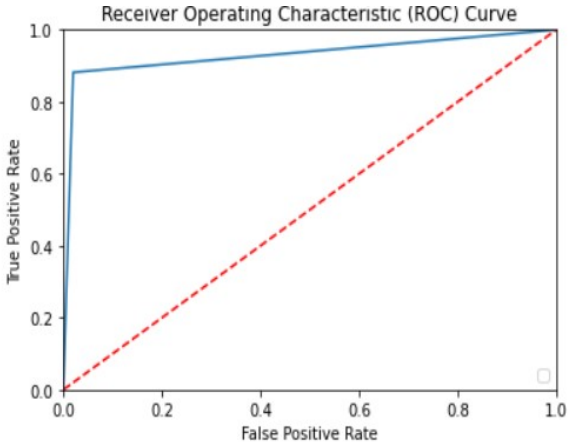

```
roc_auc_score:  0.9017712027186762
```

Fig. 3. Random Forest with Adaboost on *The Fake Project* Dataset



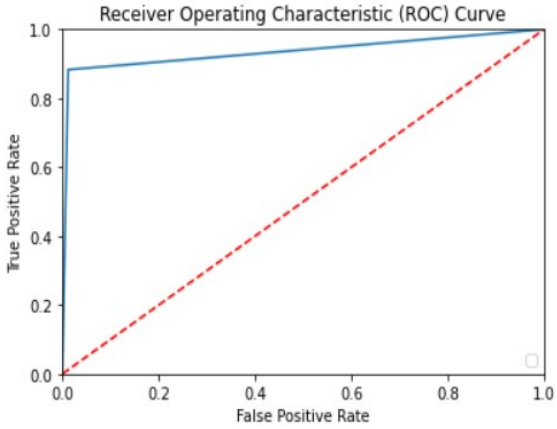

```
roc_auc_score:  0.9415843465788448
```

Fig. 5. LSTM based GAN on *The Fake Project* Dataset




```
roc_auc_score:  0.9308987002171926
```

Fig. 4. CNN based GAN on *The Fake Project* Dataset




```
roc_auc_score:  0.935482204237915
```

Fig. 6. GRU based GAN on *The Fake Project* Dataset

IV. Experiment

In this section, experimentation have been conducted to evaluate the performance of the proposed model by considering *The Fake Project* dataset [15] which contains an entirely new breed of social bots (social spambots#1, social spambots #2 and social spambots#3, traditional spambots#1, traditional spambots#2, traditional spambots#3, traditional spambots#4) and human users (genuine accounts). All these subsets of data combined together account for 7080 records of bots and 3474 records of humans with 43 features for each category. A group-wise partition of data is shown in Table II.

A. Experimental Results

Table I shows the performance evaluation of various models with and without Generative Adversarial Networks (GAN) in terms of recall, precision and accuracy on on *The Fake Project* Dataset. From Table I, it can be observed that the proposed models GAN with LSTM and GAN with GRU obtains better accuracy when compared to Random Forest with Adaboost classifier and CNN based GAN. The reason is that the proposed models solve imbalanced distribution of bots in Twitter network.

Fig. 3, Fig. 4, Fig. 5 and Fig. 6 illustrate the performance evaluation (in terms of ROC curve) of Random Forest with Adaboost classifier, CNN based GAN, LSTM based GAN and GRU based GAN, respectively on *The Fake Project* Dataset. The proposed models namely, LSTM based GAN and GRU based GAN achieves better ROC curve value when compared to Random Forest with Adaboost classifier and CNN based GAN. This is due to the fact that data augmentation is applied to generate new fake samples of data from the existing data. Further, the discriminator penalizes itself for misclassification problem by considering both residuals and discriminator results,.

V. Conclusion

In this paper, generative adversarial networks with two different recurrent neural networks models (namely long short term memory (LSTM) and gated recurrent unit (GRU)) has been presented for the identification of bots. Further, we consider user based features (such as statuses count, followers count friends count, favourites count and listed count) to analyze the behavior of online social networking user accounts. Experimentation has been conducted on *The Fake Project* dataset for evaluating the performance of the proposed modelsin terms of recall, precision, accuracy and ROC curve. Therefore, by considering both residuals and discriminator results, we have obtained on an average of 97% accuracy on precision.

References

[1] M. Kantepe and M. C. Ganiz, "Preprocessing framework for twitter bot detection," in *2017 International conference on computer science and engineering (ubmk)*. IEEE, 2017, pp. 630–634.

[2] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on twitter," *Computers & Security*, vol. 91, p. 101715, 2020.

[3] G. Lingam, R. R. Rout, D. V. Somayajulu, and S. K. Das, "Social botnet community detection: a novel approach based on behavioral similarity in twitter network using deep learning," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, 2020, pp. 708–718.

[4] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "Rtbust: Exploiting temporal patterns for botnet detection on twitter," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 183–192.

[5] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 274–277.

[6] A. Gnanasekar, S. L. R. A. Mariam, K. Deepika *et al.*, "Detecting spam bots on social networks using supervised learning," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 10 062–10 068, 2021.

[7] J. Liu, Y. Tian, R. Zhang, Y. Sun, and C. Wang, "A two-stage generative adversarial networks with semantic content constraints for adversarial example generation," *IEEE Access*, vol. 8, pp. 205 766–205 777, 2020.

[8] R. R. Rout, G. Lingam, and D. V. Somayajulu, "Detection of malicious social bots using learning automata with url features in twitter network," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 1004–1018, 2020.

[9] B. Wu, L. Liu, Y. Yang, K. Zheng, and X. Wang, "Using improved conditional generative adversarial networks to detect social bots on twitter," *IEEE Access*, vol. 8, pp. 36 664–36 680, 2020.

[10] X. Wang, Q. Zheng, K. Zheng, Y. Sui, S. Cao, and Y. Shi, "Detecting social media bots with variational autoencoder and k-nearest neighbor," *Applied Sciences*, vol. 11, no. 12, p. 5482, 2021.

[11] J. Zhang, R. Zhang, Y. Zhang, and G. Yan, "The rise of social botnets: Attacks and countermeasures," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 6, pp. 1068–1082, 2016.

[12] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Information Sciences*, vol. 467, pp. 312–322, 2018.

[13] G. Lingam, R. R. Rout, and D. Somayajulu, "Detection of social botnet using a trust model based on spam content in twitter network," in *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2018, pp. 280–285.

[14] P. Shi, Z. Zhang, and K.-K. R. Choo, "Detecting malicious social bots based on clickstream sequences," *IEEE Access*, vol. 7, pp. 28 855–28 862, 2019.

[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: efficient detection of fake twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.