JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 02, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

DETECTION OF MALICIOUS SOCIAL BOTS USING LEARNING AUTOMATA WITH URL FEATURES IN TWITTER NETWORK

[1]B. Satish, [2]S. Preethi, [3]T. Deekshitha, [4]T. Kusuma Sai Laxmi
[1]Assistant Professor, [2,3&4]UG Scholar
Department of ECE, Malla Reddy Engineering College for Women, Hyderabad

ABSTRACT

Undoubtedly, social media, such as Facebook and Twitter, constitute a major part of our everyday life due to the incredible possibilities they offer to their users. However, Twitter and generally online social networks (OSNs) are increasingly used by automated accounts, widely known as bots, due to their immense popularity across a wide range of user categories. Their main purpose is the dissemination of fake news, the promotion of specific ideas and products, the manipulation of the stock market and even the diffusion of sexually explicit material. Therefore, the early detection of bots in social media is quite essential. In this paper, two methods are introduced targeting this that are mainly based on Natural Language Processing (NLP) to distinguish legitimate users from bots. In the first method, a feature extraction approach is proposed for identifying accounts posting automated messages. After applying feature selection techniques and dealing with imbalanced datasets, the subset of features selected is fed in machine learning algorithms. In the second method, a deep learning architecture is proposed to identify whether tweets have been posted by real users or generated by bots. To the best of the authors' knowledge, there is no prior work on using an attention mechanism for identifying bots. The introduced approaches have been evaluated over a series of experiments using two large real Twitter datasets and demonstrate valuable advantages over other existing techniques targeting the identification of malicious users in social media

INTRODUCTION

Nowadays, online social networks (OSNs) have become immensely popular among users of various categories, as they can share news, opinions, organize events, collaborate or even meet new people. Twitter is a microblogging platform being used by an increasing population of users of different age groups over the last decade. People post tweets and interact with other users as well. More specifically, they can follow (following/friends) their favourite politicians, celebrities, athletes, friends and get followed by them (followers). Furthermore, Twitter generates a list of topics being discussed every day, the so called trending topics. Thus, users can get informed about the hot topics of discussion on a daily basis. However, automated accounts take advantage of these services for malicious purposes. These automated accounts, often called bots (a.k.a sybil accounts), post tweets with malicious/fake content, in order to manipulate the public opinion, sway the political discussion, promote specific ideas, products or services and spread rumours. They can also be used as fake followers, so as to increase the popularity and the reputation of a user. By posting tweets quite often, they influence measures

Submitted: 22/04/2023 **Accepted**: 10/05/2023

including the trending topics. As a consequence, legitimate users cannot distinguish between real trending topics and fake ones In this paper, two methods are proposed that aim to distinguish real Twitter users from bots, both at account and tweet level. The main contributions of this article are two-fold and can be summarized as follows. First, in contrast to existing research initiatives extracting a limited number of features, the proposed approach is based on the extraction of a high number of features (71 in total) per user to detect bots in Twitter. Moreover, several feature selection techniques have been applied for discarding redundant and irrelevant features. Second, a deep learning architecture is proposed that integrates an attention mechanism at the top of the BiLSTM layer, aiming to detect tweets generated by legitimate users or bots. To the best of the authors' knowledge, the attention mechanism has never been used for detecting tweets belonging to real users or automated accounts. Finally, it is presented that the proposed architecture outperforms earlier works, as indicated by the extended experiments conducted.

RELATED WORK Andriotis & Takasu [2] employed several machine learning algorithms to assess different sets of features, including metadata (number of statuses/followers/friends/favourites), content (retweet/hashtag/mention/URL ratio), sentiment features and LDA topics. For extracting LDA topics as a feature vector, they adopted the methodology proposed by Liu et al. [3]. They aggregated tweets from each account creating one document per user. Then they trained the LDA model extracting 200 topic probabilities for each document/user. Finally, they estimated two feature vectors, namely Global Outlier Standard Score (GOSS) and Local Outlier Standard Score (LOSS). Machine learning algorithms, including k-Nearest Neighbours (k-NN), Decision Tree (CART), Gaussian Naive Bayes (NB), Support Vector Machine (SVM), Random Forest & AdaBoost, were trained with different sets of features. The authors estimated the performance of their algorithms through 10-fold cross-validation using the evaluation metric F1-score. AdaBoost achieved the best F1-score, equal to 0.95, being trained with all the features. Another similar approach is presented by Davis et al. [4] that proposed BotOrNot's classification system, which generates more than 1000 features. These features can be grouped into six main classes: network, user, friends, temporal, content & sentiment. Random Forest Classifier is trained using the corresponded features. Likewise, Kaubiyal & Jain [5] built a feature vector per user to categorize Twitter users into real users and bots. This feature vector consisted of account-based, tweet-based, ownership detail & URL-based features. The authors employed several machine learning algorithms, including Logistic Regression, SVM & Random Forest for evaluating the proposed approach. A straightforward approach is proposed by [6] that exploited a minimal and simple enough set of features, eight in total, based on users' characteristics. They performed 20-fold cross-validation training several machine learning classifiers, including Logistic Regression, SVM & Random Forest. They, also, created an online browser plugin, so that each user may be informed if a given user id belongs to a genuine or automated account. In [7], the authors applied Random Forests, Decision Trees and Bayesian Networks as classification models for bots detection in Twitter. Furthermore, they employed a set of features, six of which have not been used before, that can be grouped into four main categories: metadata, content, interaction & community. For dealing with the imbalanced dataset, they applied SMOTE, which constitutes an oversampling technique. Another approach that also introduces new features that had not been used before is proposed by Amleshwaram et al. [8]. They identified 15 new features and employed four machine learning classifiers for detecting spam tweets. As mentioned by the authors, these features exploit the behavioural-entropy, profile characteristics, bait analysis, and the community property observed for modern spammers. The methodology presented in [9] extracts a set of features per user, that can be grouped into two main categories: account & content-based features. In order to deal with the imbalanced dataset, they applied the SMOTE+ENN technique. They ranked the proposed features using ANOVA F-value. Best evaluation results were obtained when they used all the features as input to the AdaBoost classifier. Furthermore, Ashour et al. [10] used character n-grams to detect spam tweets rather than detecting spammers. More specifically, they studied two representations of the character n-gram features, namely Term Frequency (tf) & Term frequency-Inverse document frequency (tf-idf). They implemented multiple classifiers, including Support Vector Machines, Random Forest, Logistic Regression. For conducting their experiments, they used a publicly available dataset, where they sampled one tweet for each user. On the other hand, Wang et al. [11] designed a framework by using n-grams in conjunction with further sets of features for distinguishing tweets posted by legitimate users from those posted by spammers. Finally, the feature vector per tweet consisted of user metadata (information about the user who posted the specific tweet), sentiment, content, n-grams. As far as n-grams are concerned, they tried uni+bi-gram or bi+tri-gram with binary term-frequency (tf) and tf-idf (i.e. Term Frequency times Inverse Document Frequency) techniques. For conducting their experiments, they used two publicly available datasets. They trained a Random Forest Classifier with different sets of features and evaluated their experiments using Recall, Precision & F-Measure. A more sophisticated approach aiming at identifying the subset of features yielding the best performance through data reduction techniques has been proposed by Khaled et al. [12]. More specifically, after extracting 16 account-based features, they applied Principal Component Analysis (PCA), Spearman’s RankOrder Correlation in conjunction with Markov Blanket technique & Wrapper feature selection method using SVM/Multiple linear regression. Then, they trained SVM, Neural Network & SVMNeural Network with each subset of the features chosen by the data reduction techniques. SVM-NN achieved the best classification accuracy among all the data reduction techniques. They concluded that PCA achieved poor results. A similar methodology based on feature selection techniques was adopted by Wald et al. [13]. At first, they extracted accountbased (followers/friends/description length), linguistic (LIWC) and count-based (number of days on which a given user posts a certain number of tweets from a certain category) features per user. Then, they applied feature selection techniques, including filter and wrapper methods, so as to find the subset of features yielding the best performance. They trained and tested the following machine learning algorithms: k-NN, Logistic Regression, Naive Bayes, Random Forest, SVM & MLP. The performance of machine learning classifiers trained with all features possible has been compared to the performance of techniques that measure the impact of each feature on the dataset in Herzallah et al. [14]. In particular, after classifying users into spammers or non-spammers through traditional machine learning algorithms, they used Information Gain, CoM and Relief-F for ranking features based on their importance towards the classification. They claimed that the best set of features comprises reputation of the account, age of the account, average time between tweets, average length of tweets and average mentions per tweets. A more advanced and quite different approach with regards to the aforementioned ones was adopted by [15,16], who followed the methodology introduced in [17]. More specifically, Pasricha & Hayes [15] modelled the behaviour of each account according to the type of tweets this account posts. After compressing the DNA sequence, they extracted the size of the DNA before and after compression as well as the compression ratio. They trained two logistic regression classifiers, using as features the compressed DNA size for the one classifier, whereas they used both the compression ratio and the size of the original uncompressed DNA as input to the other classifier. Similarly, Kosmajah & Keselj [16] after defining a set of codes, converted each tweet into a character label by using ASCII table character indexes and in this way they formed a sequence of characters for each user. Next, they extracted n-grams with n ranging from one to three and finally they applied several statistical diversity measures. They evaluated their approaches performing 10-fold cross-validation and using F1-score as the evaluation metric. They trained several machine learning classifiers, including Gaussian Naive Bayes, SVM, Logistic Regression, k-NN, Random Forest and Gradient Boosting. Best performance was achieved, when they used uni+bi+tri-gram features as input to the Random Forest Classifier.

EXISTING SYSTEM

In [11], authors annotated more than 8000 accounts and proposed a classifier which achieved a considerable level of accuracy for such set of samples. Additionally, [38] presented a model for Twitter bot detection based on a large number of metadata from the account to perform the classification. More recently, several scientific studies have incorporated more annotated samples to support this research such as [27], [44], [45] including some procedures for achieving better level of accuracy by strategically selecting a subset of training samples that better generalize the problem. In [22], a language-agnostic approach is employed to identify potential features to distinguish between human and bot accounts. The model is then trained and validated using over 8000 samples distributed in an unbalanced fashion and its performance reaches an accuracy of 98%. Moreover, authors in [32] proposed a 2D Convolutional Network model based on user-generated contents for detecting bots from human accounts including its gender (male, female account) covering both Spanish and English languages. A similar goal is explored by authors in [40], where both Word and Character N-Grams are employed as main features to perform the classification. A different manner of addressing the problem was recently proposed by authors in [5], where novel altmetrics data to investigate social networks are analysed and they are used to train a Graph Convolutional Network (GCN) which reaches over 70% of accuracy in this task. On the other hand, authors in [34] presented a novel one-class classifier to enhance Twitter bot detection without any requiring previous information about them.

DISADVANTAGES

1) The system doesn’t have A MULTILINGUAL APPROACH FOR USER ACCOUNT ENCODING VIA TRANSFORMERS.

2) The system couldn’t implement to detect the following (i) Level of activity, (ii) Level of popularity, (iii) Profile information.

PROPOSED SYSTEM

Present a multilingual approach for addressing the bot identification task in Twitter via Deep learning (DL) approaches to support end-users when checking the credibility of a certain Twitter account. To do so, several experiments were conducted using state-of-the-art Multilingual Language Models to generate an encoding of the text-based features of the user account that are later on concatenated with the rest of the metadata to build a potential input vector on top of a Dense Network denoted as Bot-DenseNet. Consequently, this paper assesses the language constraint from previous studies where the encoding of the user account only considered either the metadata information or the metadata information together with some basic semantic text features. Moreover, the Bot-DenseNet produces a low-dimensional representation of the user account which can be used for any application within the Information Retrieval (IR) framework.

ADVANTAGES

- A preprocessing stage where a multilingual input vector of the user account is generated
- A final decision system for identifying whether the account has a normal or abnormal behaviour according to existence patterns in the input vector generated during the first stage.

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

Login, Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Predicted Tweet Account Type Details, Find Tweet Account Type Ratio, Download Predicted Data Sets, View Tweet Account Type 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, Predict Tweet Account Type, View Your Profile.

CONCLUSION AND FUTURE WORK Given the proliferation of malicious accounts in social media, the early detection of these evolves into a critical challenge necessary to be addressed. More sophisticated techniques need to be exploited, so as to address the above–mentioned limitations and therefore, to mitigate bot activity in social media. In this framework, this paper introduces two novel methods that aim to detect bots in Twitter, both at account and tweet level. In the first case, feature extraction and selection techniques have been exploited to identify Twitter bots. In this respect, a thorough analysis over the features used has been presented. In the latter case, the usage of an attention mechanism led to a substantial improvement with regards to the evaluation metrics. Via series of experiments conducted, it has been demonstrated that the proposed approaches clearly outperform the related state-of-the-art results. More specifically, in the first approach, after conducting a comparison among several feature selection and dimensionality reduction techniques, the techniques that lead to efficient and stable features that boost the performance of machine learning classifiers have been identified and selected. The respective experimental results indicate that Logistic Regression appears to be the most effective feature selection technique. In terms of evaluation metrics, the usage of Logistic Regression and Random Forest adopted as for feature selection and classification respectively demonstrated 0.9906 accuracy, which outperforms the existing state-of-the-art approaches considered. Similarly, the usage of Logistic Regression and SVM for feature selection and classification respectively demonstrated 0.9977 AUROC, which also outperforms the other existing approaches investigated. With regards to the other evaluation metrics considered, i.e., recall, precision and F-measure, less significant improvements are observed. With respect to the second approach, the obtained results indicate that the proposed approach outperforms by 5.7% in average in terms of precision, by 23.6% in average in terms of recall, by 15% in average in terms of F-measure, by 7.7% in terms of accuracy, and by 14% in average in terms of AUROC. Overall, there is a single occasion where the proposed approach is outperformed by another for one of the evaluation metrics. This limited evidence is very weak to question the validity of the obtained evaluation results indicating the superior performance of the proposed approach

REFERENCES:

[1] M. Orabi, D. Mouheb, Z. Al Aghbari, I. Kamel, Detection of bots in social media: A systematic review, Inf. Process. Manage. 57 (4) (2020) 102250.

[2] P. Andriotis, A. Takasu, Emotional bots: Content-based spammer detection on social media, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–8.

[3] L. Liu, Y. Lu, Y. Luo, R. Zhang, L. Itti, J. Lu, Detecting ‘‘smart’’ spammers on social network: A topic model approach, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 45–50.

[4] C.A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, Botornot: A system to evaluate social bots, in: Proceedings of the 25th International Conference Companion on World Wide Web, 2016, pp. 273–274.

[5] J. Kaubiyal, A.K. Jain, A feature based approach to detect fake profiles in twitter, in: Proceedings of the 3rd International Conference on Big Data and Internet of Things, 2019, pp. 135–139.

[6] A. Narayanan, A. Garg, I. Arora, T. Sureka, M. Sridhar, H. Prasad, Ironsense: Towards the identification of fake user-profiles on twitter using machine learning, in: 2018 Fourteenth International Conference on Information Processing (ICINPRO), IEEE, 2018, pp. 1–7.

[7] M. Fazil, M. Abulaish, A hybrid approach for detecting automated spammers in twitter, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2707–2719.

[8] A.A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, C. Yang, Cats: Characterizing automation of twitter spammers, in: 2013 Fifth International Conference on Communication Systems and Networks (COMSNETS), IEEE, 2013, pp. 1–10.

[9] J. Knauth, Language-agnostic twitter-bot detection, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019, pp. 550–558.

[10] M. Ashour, C. Salama, M.W. El-Kharashi, Detecting spam tweets using character n-gram features, in: 2018 13th International Conference on Computer Engineering and Systems (ICCES), IEEE, 2018, pp. 190–195.