International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3537-3541
 https://dx.doi.org/10.37624/IJERT/13.11.2020.3537-3541

Identification of Bots in Social Networks based on Data Mining Technologies

Ilyas Idrisovich Ismagilov[1], Ajgul Ilshatovna Sabirova[2], Dina Vladimirovna Kataseva[3], Alexey Sergeevich Katasev[4]

[1]Doctor of Technical Sciences, Professor of the Department of Economic Theory and Econometrics, Institute of Management, Economics and Finance, Kazan (Volga Region) Federal University; Scopus ID: 6603342575; ORCID: 0000-0002-0446-8204, Kazan Federal University, iiismag@mail.ru

[2]Ph.D. in Economics, Senior Lecturer at the Department of Accounting, Analysis and Audit of the Institute of Management, Economics and Finance, Kazan (Volga Region) Federal University; Scopus ID: 56646429700; ORCID: 0000-0002-6734-740X, Kazan Federal University, aigylkinyes@mail.ru

[3]Senior Lecturer, Department of Information Security Systems, Institute of Computer Technologies and Information Security, Kazan National Research Technical University named after A.N. Tupolev-KAI; Scopus ID: 57193401954; ORCID: 0000-0001-6141-8329, Kazan National Research Technical University named after A.N. Tupolev, DVKataseva@kai.ru

[4]Doctor of Technical Sciences, Professor of the Department of Information Security Systems of the Institute of Computer Technologies and Information Security, Kazan National Research Technical University named after A.N. Tupolev-KAI; Scopus ID: 57193408902; ORCID: 0000-0002-9446-0491, Kazan National Research Technical University named after A.N. Tupolev ASKatasev@kai.ru

Abstract

This article solves the problem of intelligent models constructing and their accuracy evaluating for identifying bots in social networks. The relevance of solving this problem is noted. The construction and accuracy assessment of the neural network model, decision tree and linear regression are performed. The initial data source was Twitter social network. To collect the initial data, we used our own database, consisted of 3428 users, about half of which contained characteristic features of bots. The initial data were randomly divided into the training and testing sets, each of them included approximately 50% of the records. 15 attributes were used as the model's input parameters, in particular, the number of symbols in the username, the user's number of tweets, the number of readers, etc. The models construction and study was carried out on the Deductor analytical platform base. Each model was tested on data set consisted of 1719 records. For all models, the corresponding classification matrices were constructed and the first, second kind errors and the general model's error were calculated. In terms of minimizing these errors, the neural network model showed the best results, and the linear regression model showed the worst. This allowed us to conclude, that in order to minimize classification errors, it is advisable to use a neural network model. This indicates its effectiveness and the possibility of practical use in intelligent decision-making support systems for bots identifying in social networks.

Keywords: Bots Identification, Social Network, Neural Network, Decision Tree, Linear Regression, Decision-Making Support.

I. INTRODUCTION

Currently, social networks have become very popular among users [1]. The social networks audience is growing due to cheaper mobile devices with access to the Internet, and due to the widespread availability if this network. More than 70% of Internet users are registered in various social networks [2,3].

Social networks have long ceased to be considered as platforms only for communication and entertainment. They attract large audiences due to the accumulated data sets about their users [4,5]. Therefore, almost all medium and large companies are represented on social networks. Public persons are usually also registered in all popular blogging platforms. Such popularity attracts the attention of various attackers, creating bots on social networks [6-8].

A bot is a program that automatically performs any actions in a social network [9]: reply to messages and send them, commenting messages, setting «likes», etc. Such programs quickly created millions of user accounts masquerading as real people, and flooded the largest social networks such as Facebook, Vkontakte, Twitter, Instagram and others.

Currently, bots creation and use in social networks is not difficult [10]. There are a large number of programs for creating bots in Internet. At the same time, it is very difficult to prove that the page owner in a social network used or created bots himself.

The goals of using bots can be divided into three groups [11]:

- increasing the credibility of public figures, business organizations, media;
- products and services promotion (spam);
- phishing.

The information source popularity directly affects on its perception. Therefore, bots are often used to increase the authority of public figures, for example, politicians. The accounts popularity achieved through bots leads to real user's attraction. Bots are also used in the media to increase the citation of their publications.

Bots can be used for distribution of unfair advertising (spam)

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3537-3541
 https://dx.doi.org/10.37624/IJERT/13.11.2020.3537-3541

[12,13]. Since the majority of social network users don't expect advertising from other users, such bots are actively promoting any products and services, disguising messages as positive reviews.

The third type of bots, which is dangerous for users, is bots for phishing [14] (obtaining secret information by deceiving the user) and distribution of malicious software [15].

Thus, the problem of the bots distribution in social networks is relevant. Bots remove the difference between useful and useless information, clog the information field, pushing users away from social networks, reducing the effectiveness of fair advertising, and harming organizations that own social networks and their customers. Bots can defame other people's names and company brands, causing reputational damage. Bots also bring fame to the incompetent media. In addition, bots can lead to defeat of computers with malware and the loss of confidential information.

Therefore, the effective models development for bots identifying in social networks is relevant [16].

II. METHODS

In this work, a neural network, a decision tree, and linear regression are used as tools for constructing bots identification models in social networks [17-19]. The choice of these methods is due to their high efficiency in solving problems of diagnostics, pattern recognition, objects classification, and their condition assessment [20-23].

The neural networks' use for classification problem solving consists of affiliation indicating of input image, represented by a feature vector, to one or more predefined classes [24,25]. This principle is base for its application to the bots identification in social networks. The solution of this task with the help of a neural network consists of user's assigning to one of two categories («bot» / «not bot»).

Unlike neural networks, decision trees are a way of representing rules in a hierarchical, sequential structure, where each object corresponds to a single node that provides a solution. The goal of the whole process of constructing a decision tree is to create a model, by which it would be possible to classify cases and decide what values the objective function can take, having several variables as the input.

The linear regression is designed to obtain a forecast of continuous numerical variables and to solve the classification task. Regression is the conditional mathematical expectation of a continuous dependent (output) variable with the observed values of independent (input) variables. The classification task in linear regression is the task of assigning an object to a particular class. To determine the class membership, a certain set of object parameters was used. These parameters were represented as a numerical vector of fixed dimension. By analyzing these parameters values, a decision of network user classification was made.

At the first stage of constructing bots identification models, features for classifying social network users were selected. Since users' pages are the collection of data in electronic form, it is necessary to obtain from the initial information a clear set of numerical parameters characterizing the bots, the values of which then will be included in the training and testing sets. The most important and informative features are presented in Table 1.

Table 1. Social network' users classification features

No.	Feature name	Feature description
1	Number of symbols in the name	Number of symbols in the name (nickname)
2	Number of tweets	Number of tweets in user's page
3	Number of users to read	The number of user's page which the user has subscribed
4	Number of readers	The number of users who are subscribed to the user's page
5	Favorite	Bookmarked tweets
6	Frequency of occurrence of retweets	Frequency of messages that were copied from other users
7	Frequency of user's posts occurrence	Frequency of occurrence of messages that were written by the user on his page
8	Hashtag frequency	Hashtag frequency
9	Response frequency	Frequency of responses to messages to other users
10	Frequency of advertising occurrence	Frequency of advertising messages occurrence on a user's page
11	The presence of a field with information about user	The presence of a field with information about user on the user's page
12	Links to additional sites	Links to additional sites
13	Location	The presence of a field with information about user's location
14	Registration date	Number of days since user registration

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3537-3541
 https://dx.doi.org/10.37624/IJERT/13.11.2020.3537-3541

The information content of these features was evaluated experimentally using mathematical statistics methods. Then the data set from various sources was created. The obtained data were presented in the form of a table, where each row corresponded to individual users, and each column corresponded to the characteristics for their classification.

The Twitter social network [26] was a data source with its own database of 3428 users, about half of which contained characteristic features of bots. The initial data were randomly divided into the training and testing sets, each of which included approximately 50% of the records.

III. RESULTS AND DISCUSSION

For social networks bots identifying models construction, the Deductor analytical platform was used. It includes all the necessary tools for data importing, a neural network model, decision trees, linear regression constructing and their effectiveness evaluating. The construction of these models was carried out experimentally by repeatedly training them and selecting the optimal parameter values at which the maximum classification accuracy was achieved.

From the point of view of architectural aspects, the neural network with two hidden layers of 10 neurons in each layer turned out to be optimal. For the neural network training the *Back Propogation* algorithm was used. The criterion for stopping the training was the classification error minimization in the testing set used in the training process.

The constructed decision tree included the following set of optimal parameters:

- minimum number of examples in the node: 3;

- construct a tree with more reliable rules to the detriment of its compactness;

- cut off tree nodes at a confidence level of 25%.

The first two parameters are responsible for stopping the decision tree construction. The third parameter is used to cut off tree nodes: a lower level of trust results leads to more cut nodes.

When the linear regression model constructing, the method of variables selecting was choosen, as well as the probability of including and removing variables from the model. Full inclusion (*Enter*) was chosen as the optimal method for variables selecting, the probability of including a variable in the model was 0.05, and the probability of deleting a variable from the model was 0.1.

The results of the constructed models accuracy assessing on the training set are presented in the table 2 [27].

Table 2. The models results on the training data set

Actual values	Classified by the model						
	0			1			Total
	NN	DT	LR	NN	DT	LR	
0	833	820	784	23	36	72	856
1	18	27	37	835	826	816	853
Total	851	847	821	858	862	888	1709

In this table we used the following notation: NN - neural network, DT - decision tree, LR - linear regression. Moreover, the number «1» means «bot», and the number «0» means «not a bot». Table 3 presents the results of the constructed models accuracy evaluating on a testing data set.

Table 3. Models results on a testing data set

Actual values	Classified by the model						
	0			1			Total
	NN	DT	LR	NN	DT	LR	
0	840	832	791	21	29	70	861
1	19	32	38	839	826	820	858
Total	859	864	829	860	855	890	1719

Based on the data presented in the table, the 1st, 2nd kind errors [28], and the general error of the models were calculated (see Table 4).