How to identify bots using mathematical modelling and IT tools

January 2019
1. Features of bots

Definitely the largest part of the Internet traffic generated by the bots comes from bots, whose task is to make their behavior similar to human behavior. Thanks to this, they can take on made-up identity and influence opinions and decisions of real people. The construction of algorithms imitating human thinking and actions is currently at a very high level and is based mainly on the achievements of the so-called artificial intelligence (AI). This means in short that the artificial intelligence can be used to specify any technique that allows a computer program to imitate human intelligence using logical rules, decision trees, machine learning, or other research techniques. A schematic overview of the ongoing work on artificial intelligence is presented in Figure 1.

![Figure 1. Schematic overview of the areas in which works on artificial intelligence are involved. Source: The Ember Group 2017.](image)

The set of the categories of good and bad bots, the diversity of examples and the universality of applications, the multifaceted nature of threats, and even the multilayer and complexity of individual sources of development of particular types of bots - all this affects the inability to present a coherent approach to bots and their identification. This task requires a *de facto* dedicated approach not only
to each type or category of bots, but also to any specific problem of the origin and purpose of their existence. Each of the proposed solutions must take into account the individual character of the considered issue. In another way, we will have to fight brute force or DDoS hacking, and in a different manner detect fraud attempts on the financial market, or count profitability statistics related to personalized ads displayed in real time. Each of those tasks requires a different approach and no solutions can be offered that would be equally effective in each of the mentioned examples. Thus, the review of possible and even existing approaches to identification, not to mention the ways of counteracting actions performed by some bots, is virtually impossible in a single study. Therefore, in subsequent chapters, a more detailed presentation of the aspects was provided only for exemplary applications. However, for the sake of coherence of the description, in this chapter the attempt to present in some sense a part common to the bots problem covering the features of their operation was undertaken. In the next chapter, the focus was on listing and brief overview of those methods or algorithms that potentially have or may have the greatest importance in solving identification problems, filtering or protecting against generally understood adverse consequences of the bots.

1.1. Features that allow the identification

Although each of the bots has its own specificity, some common features can be indicated at least for a few of automatic programs that replace people, if not for all of them. Intuitively, the task of the bot is to perform the work previously performed by man. The difference is that, in general, the bot performs it faster, more efficiently, longer and flawlessly. Therefore, the basic issue is that the computer program is supposed to imitate human behavior or activities performed thereby, but it does it better, because should it do it much worse, it would pay more to leave a given job to people. The man’s imitation, no matter how perfect it was, should, at least once in a while, arouse suspicion, due to the kind of unpredictability and, consequently, the non-determinism of human activities. Therefore, exploration of the places of potential differences between machine and human behavior should allow to distinguish them. The following are considered to be quite ordinary, used in practice methods of distinguishing bots from people based on simple metrics resulting from common-sense (human) limitations.
• Long or even uninterrupted operation

People must sleep, eat, rest, ensure the fulfillment of basic physiological needs, etc. Therefore, it is not physically possible for a man to perform any activity for a very long time. Of course, the problem is to determine from which moment one can talk about too long time. One can mention sporadic cases of, e.g., participating in a non-stop game for several or several dozen hours. There are even fatal cases like that of Seungseob Lee in 2005, who played 50 hours continuously in StarCraft, dying of a heart attack\(^1\). However, this should be considered as an (extreme) outlier from the statistical point of view. Statistically speaking, those human limitations can be one of the basic methods of distinguishing a man from a bot that is not burdened with such (biologically) restrictive determinants.

Therefore, in order to identify the bots, one can measure the user’s activity at different times of the day and night (preferably continuously). Exceeding reasonable limits, when it comes to the total time of activity, indicates the activity of the bot, while the sum of times alone may not be a sufficient criterion, because one can imagine a bot whose one-time action is short, but the number of those activities over the day has a surprising intensity. Too short breaks between individual activities may also indicate that we are not dealing with a human being.

• Extraordinary performance or effectiveness

One of the basic limitations of man is the rapidity of his action. Even when compared to other biological organisms, human physical parameters are much weaker. It is, among other things, about strength, speed or endurance. The predominance of computers in comparison to people, is even more explicit than in the case of animals. This can be used to distinguish bots from people. It is enough to measure the number of actions taken and the times in which they occurred. There are quite a lot of examples of such measures. It is possible to mention the number of site

\(^1\) [https://starcraft.fandom.com/wiki/Lee_Seung_Seop](https://starcraft.fandom.com/wiki/Lee_Seung_Seop) (last accessed date: 06/01/2019).
views or visits on different sites. It seems reasonable to assume that a person cannot view tens of thousands or millions of websites (at least in a short time). The statistics of visits are just one of the possibilities. An alternative is, e.g., measuring the number of actions performed on the websites. The actions mean any activity performed within a given website. Therefore, it can be clicking on a link, button or banner, as well as site scrolling to the right place\(^2\). It can also be the number of video playbacks or even the number of hovering over the banner (without clicking). In each of those cases, exceeding the set limit is a premise that we are dealing with an automaton, not a living person. Physically it is impossible (or unlikely) by the human to achieve values that are easily achievable for computer programs.

It is also important to specify that the limits mentioned depend on the specifics of the issue and the preferences of the person to identify the bots. Setting the limits too high will cause that some bots will be considered people. On the other hand, setting limits too low will result in the rejection of some users who are actually human. However, the choice of the optimal solution is purely subjective and depends on the purpose of a given person.

- Incredible rapidity of action

An analogous parameter to the abnormal performance mentioned above is the unreliable or unlikely rapidity of action. It is obviously extremely desirable in many cases, including auction algorithms or searching for opportunities for immediate earnings resulting generally from human errors. An example may be the setting, by mistake, of too low the price of the sold product or the emergence of arbitrage opportunities on the stock exchange. The rapidity of action is also critical in RTB (Real Time Bidding) systems, which will be discussed later in the study.

Some of those situations undoubtedly require automation. However, in many cases we expect the user to be a real person. Suspicious actions, in

---

\(^2\) Measurements related to scrolling are often to insert pixels into the appropriate places on the site, whose activation means that the user has reached the indicated place on the site. For example, getting to 3/8 sites is identified as activation of scroll_3_8 (or any other) pixel, while reading 6/8 sites corresponds to scroll_6_8 pixel.
this case, should be a surprisingly fast response time to any event or extremely fast execution of any activities. And so, it is expected that a person would like to become familiar with the content of the site for a while before clicking on an advertisement or going down to it. Monitoring the time that passes from the moment of entering the website to the time of performing specific actions, or the time to fill a form, or even the speed of entering text into a field, is one of the key measures on the basis of which one can distinguish a living person from an artificial machine.

• Surprising randomness

Randomness is one of the most surprising factors for identifying the bots. However, it is not trivial in any respect. On the one hand, we are not dealing with “real” randomness on computers. We speak of pseudo-randomness in this context. On the other hand, this pseudo-randomness is much “more random” than human behavior. Contrary to appearances, the human perception of randomness is very subjective. For example, it was empirically proven that people tend to select odd numbers more often than even numbers. Also, prime numbers are selected statistically significantly more often than complex numbers.

People’s behaviors are certainly not random, and in many cases, they fit easily-defined patterns. Some people’s activities can be predicted with a surprisingly high probability of over 90%. Meanwhile, the operation of computer programs is, in principle, deterministic with the possible addition of random elements. It is possible to distinguish those two approaches provided that one has enough data at our disposal. Let’s examine two examples. Bots, one of the activities of which is to visit websites (for any purpose), include visits programmed in them according to a fixed or random schedule. Too much repeatability of the visiting scheme indicates the lack of interference of a living person, and thus the activity of the bot. If, on the other hand, the randomness is inscribed in the algorithm, then based on


many data, due to the pseudo-randomness, we most likely get the density of one of the ordinary probability distributions (usually normal or uniform distribution). In the case of people's activities, this randomness will be rather distorted. In conclusion, both in cases when we are, and when we are not dealing with a random algorithm, one can make an estimate related to the unknown identity of the user by describing it as a bot or a human.

Another example related to the issue of randomness is the generation of usernames (logins) or email addresses. Man, usually chooses his name, nickname or some abbreviation related to personal data, adding, e.g., the year of birth (due to the unavailability of a preferred, simpler name). Following those noticed regularities, bots generate passwords e.g. by randomly connecting names with a string of numbers or by selecting a completely random string of characters. The names generated in this way are doubtful, because who chooses a name that cannot be meaningfully pronounced (e.g. “rtfgdjhbzc”) or in what year would Kamil, known as Kamil6743, be actually born? Of course, this is not a rule, but it is certainly a premise to treat someone, with such a log in, as a bot.

• Abnormal variability

Most people use computers as tools and get used to certain hardware and software configurations. High aversion to changes, or the desire for some kind of stabilization, or maybe simply lack of time or skills - for whatever reason, people rather rarely change the settings of the equipment they use. Therefore, the users whose operating systems or browsers change from day to day are extremely suspicious and based on the frequency of changes in the configuration of the hardware or software used, a bot can be identified.

However, too high changeability is also an indication in other areas. For example, there are people who are fluent in two languages, and insert posts on Twitter or Facebook in both languages. However, they are generally the inhabitants of countries in which those languages are official (or at least very popular), such as Canada or Switzerland. However, the combination of, e.g., Hungarian and Arabic, may be suspicious, especially if the language in
which the posts are written has changed suddenly and for no apparent reason. And if in addition such language changes occur very often, then we can be sure that we are dealing with a bot.

- **Specificity of human behavior**

  Those regularities in human behavior allow to establish a pattern, the significant deviations from which draw suspicions about the actual identity of the user. Those regularities can be of various types. Some of them are related to biology and physiology, others are based on research in the field of psychology or sociology. Most importantly, they can be used to identify the bots.

  One of the previously mentioned human limitations is of course his speed and effectiveness. Others include relatively poor parallelisation of works. By measuring the number of simultaneously performed activities, or even open and active connections, we are able to determine that a human would not be able to do it. For example, such suspicious activities are watching many video materials at the same time, either scrolling through multiple sites at once or simultaneously filling out several forms. Therefore, the parallelization is one of the factors that allows users to be classified into the bots or people category.

  However, there may be more such regularities, but their determination may not be an easy matter and is often a real research challenge. Potentially, the behavior may be affected, e.g., by the day rhythm in a given country (the existence of siesta or not), the attitude to holidays, but also the statistics of sites closed before reading them, and the specificity of moving the mouse.

  However, all the limits listed above should be adapted to the specifics of the activity being examined. Moreover, sometimes one of those measures may not be sufficient to make an unambiguous classification. Let's examine the following hypothetical situation. The given user is characterized by the above-standard number of websites visited. However, there are mechanisms built into browsers
that allow the user to simultaneously open all selected articles in separate tabs. This means, on the one hand, the help of an automation, but the action itself is performed by a man and it is difficult to classify it as a bot operation. Nevertheless, the situation where an action is performed on each of those many sites in a short time, seems to be a sufficient argument to classify this user as a bot. In conclusion, neither the number of sites, nor a quick click or scrolling on the selected site seem to be enough in this case, but the combination of those two measures is perfectly possible. Therefore, it seems important to combine the selected metrics in one coherent classification strategy. You can use for this, e.g., the linear regression approach or classification trees discussed in the next chapter.

You can also create aggregated measures such as viewability, defined, e.g., as an advertisement whose half of the surface was visible on the screen for a minimum of one second (for a display advertising) and half of the surface of the video advertising player was visible for two continuous seconds during the advertising playback. Therefore, this measure includes two following parameters - at to what part of the advertising space and how long it was visible. A more detailed analysis of this measure was made, among others in the document published by the IAB.

1.2. The issue of classification

Despite such a large variety of bots, their goals, participation in Internet traffic, as well as the level of complexity of their software, the very task of identifying them comes down to the standard classification issue, but we are usually talking about the simplest case of assigning to one of the following two sets: people or bots. There are many classification methods, and their effectiveness depends primarily on the available information about the object that we want to classify. The following methods are worth mentioning:

- k-nearest neighbors algorithm,
- Bayesian networks,
- logistic regression,

• decision trees,
• neural networks,
• random forests.

Most of them will be described together with examples in the next chapter, but for more information it is worth referring to numerous literature publications. The approach from a practical point of view is presented, e.g., in Charu C. Aggarwal’s book Data Classification: Algorithms and Applications⁶.

The issue of classification is the basis of the machine learning, one of the main sections of which is the supervised learning⁷. It is based on a set of predefined features designated for training data. Optimization of settings related to the indicated features on the training set, allows also a good fit of new, unprecedented cases (test set). There is also the unsupervised learning called clustering, in which computers find regularities or common features, based on which clustering of data occurs, while the clusters are not predetermined (in contrast to the supervised learning). The key to the problem of classification is the selection of the appropriate classifier, which maximizes the chances that the conclusions drawn from the analysis of the training set will be transferable to observations outside that set.

There are many examples of classification issues in the modern world. Below are a few selected ones:

• The so-called customer scoring is calculated by the bank for people who apply for a loan. The customer profile is analyzed using the machine learning algorithms that assess credit risk. People with a high-risk ratio are considered bad borrowers and are not likely to be granted a loan.

• The list of videos proposed for a given user on the platforms like Filmweb or Netflix is built based on the history of ratings, viewing and classification algorithms, also more commonly known as the recommendation systems.

---


• The proposals for extending contracts with telephone operators received before the end of the validity of the current contract are the result of classification algorithms aimed at the customer retention.

In the following part of this chapter, examples of interesting and, above all, diversified use of the classification issues in the topic of Internet bots will be described in more detail. We will refer to this example also later in the study.

1.2.1. Spam filtering

One of the frequently provided examples of the classification issues is the detection of spam among email messages. Spam is a phenomenon that almost everyone who uses the Internet will meet. Due to the scale of the phenomenon (cf. information on spamming from the previous chapter), each modern email has a built-in anti-spam filter or the ability to create some kind of rules related to incoming emails (e.g. white list, black list). Anti-spam filters are getting better, among others due to the possibility of learning on larger and more representative datasets, as well as progress in the field of the machine learning.

It is relatively easy to find attempts to approach this topic by many programmers and data analysts. On the one hand, this problem is considered a canon of the machine learning, and on the other hand, of the access to many free datasets on this issue, including to the database named “Enron Email Dataset”\(^8\), which contains 33716 emails belonging to over 150 employees (mainly management board members) of Enron. Originally, they were secured during the investigation by the Federal Energy Regulatory Commission, and after making them public, “CALO Project” organization\(^9\) dealt with their processing. The emails were deprived of attachments, and some of them were removed at the request of employees. The improperly saved email addresses have been changed to user@enron.com, if the recipient was mentioned by name and to no_address@enron.com otherwise. All messages are in English.

Based on the selected 5172 emails, a test of the effectiveness of the three different classifiers was performed. Methods have been chosen for this purpose,

---

\(^8\) https://www.cs.cmu.edu/~./enron/ (last accessed date: 06/01/2019).

\(^9\) http://www.ai.sri.com/project/CALO (last accessed date: 06/01/2019).
which could potentially be or are known to be used in such issues. They included a naive Bayes classifier, a decision tree and a maximum entropy method. Some of those methods will be described in more detail in the next chapter. Here we will focus on presenting the results of a simple study that are included in Table 1. It can be seen that it is possible to achieve over 90% effectiveness using only the content of emails. This effectiveness can be improved by using, inter alia, better data transformations, a larger set of email features (e.g. IP numbers, sender data) or more sophisticated classifiers. However, the key point is that the identification of spam sent by the bots is very real, and its operation can be observed in practice when using your own email program.

Table 1. Comparison of results of classification of three classifiers on the training set (80% of data) and the test set (20% of data).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Relevance of classification on the training set</th>
<th>Accuracy of the classification on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes classifier</td>
<td>95.91%</td>
<td>94.69%</td>
</tr>
<tr>
<td>Decision tree</td>
<td>99.59%</td>
<td>92.66%</td>
</tr>
<tr>
<td>Maximum entropy method</td>
<td>70.63%</td>
<td>72.46%</td>
</tr>
</tbody>
</table>

Source: own study based on the master's thesis of Artur Dzięcioł.

1.3. Problem of forecasting

The issue of prediction seems to be, apart from the classification, the most important task whose quick and error-prone solution allows to improve the effectiveness of many activities related to the issue of Internet bots. The results of forecasting can be used, among others by transactional bots (exchange rate forecasts, auction platforms) or information bots (prediction of weather changes, car traffic intensity, recommendation).

The forecasting methodology is very extensive but quite well described in the literature. The frequently used forecasting methods include time series forecasting (exponential smoothing model, moving average, decomposition
methods, statistical analysis, etc.), quantitative forecasting (continuous variable models, discrete variable models) and expert methods.

An example of the Real Time Bidding system and the forecast of Internet traffic volume from a wide range of predictive algorithm applications will be presented.

1.3.1. Real Time Bidding

Real Time Bidding (RTB) is a method of displaying online ads in real time. Due to the strict time limit (usually less than 120 ms) the method is reserved only for automated systems using highly specialized algorithms and efficient computers. Usually, the company’s profit is related to the activity performed by the user after viewing the advertisement on the website. Displaying the advertising without any interaction with the user means a loss for the company. Therefore, the mathematical models are created in order to select the users who are most likely to perform an activity (e.g., click on an ad, fill in the form, subscribe to the newsletter).

RTB is a model example of what Big Data issues entail. By definition, the Big Data is characterized by:

- large amount of data (currently at least millions of objects),
- large variety of data,
- variability, i.e. high speed of data receiving.

For the reasons mentioned above, it is not possible to distinguish the users manually. Dedicated, automatic traffic filters are the only solution to the problem of cheaters detecting.

Displaying one advertising is usually cheap (counted in micro dollars), but the number of such advertising is counted in tens of billions a day. Thus, without a control mechanism, the company could spend the entire budget in a few minutes. However, using the predictive algorithms, one can ensure a cost control in the RTB system related to the online bidding traffic. Hourly spending limits are

---

10 Bernardelli M. (2017), Predicting hourly Internet traffic in the RTB system - panel approach, “Roczniki” Kolegium Analiz Ekonomicznych SGH, No. 47, Oficyna Wydawnicza SGH, Warszawa, p. 5-17
determined based on historical data. This approach allows to spread the costs throughout the day instead of spending them at the beginning of the day. In order to improve the accuracy, the proposed method considers the panel econometric model (see subchapter 3.7), in which hours are the panels. The results are evaluated on the basis of comparative tests between the panel model (fixed effects model) and the time series model (least squares estimator). It turns out (see Figure 2) that in most cases the panel method gives more accurate forecasts ordinary.

![Figure 2. Percentage of costs per hour for an example day with two forecasts indicated: a panel fixed effects estimator (solid line) and a ordinary least squares method (dashed line).](source: Bernardelli M. (2017))
2. Review of algorithms

There are many methods, techniques and algorithms used in the machine learning, artificial intelligence, econometrics or statistical modeling. However, some of those methods better meet the needs related to the identification of bots. Due to the variety of applications, it is impossible to list all the available approaches, hence this chapter focuses on those methods that are known to be used in the bot-related issues, or the applications are not known, but have a lot of potential, estimated based on the other known problems. At the same time, it is worth noting that the vast majority of those methods are applied in the issues mentioned in the previous chapter, i.e., the classification and prediction.

Descriptions of the methods are not complete, in the sense of reviewing theoretical assumptions, derivation of models and interpretations. Full descriptions are in the books on this subject, some of which are provided by specific methods. In this chapter, we focus rather on presenting the concept of the algorithm and assessing its suitability in selected cases. For the sake of simplicity, the descriptions are modeled primarily on online materials\(^\text{11}\) which are often more condensed, although not entirely accurate.

2.1. Linear Regression

The linear regression is undoubtedly the most common way of modeling relationships in any area of science and practice. It is considered as one of the simpler algorithms and the numerically stable methods for determining the parameters of the linear model are known, and the obtained results are relatively easy to interpret. It should be remembered that there is a whole econometric theory, hidden under the abbreviation of the Ordinary Least Squares Method, according to which a number of assumptions\(^\text{12}\) should be checked and appropriate statistical tests performed. Only their fulfillment allows for the interpretation of

---


\(^{12}\) Gauss-Markov theorem.
the estimated parameters of the model. You can read more about the linear regression, e.g., in Wooldridge’s publication\textsuperscript{13}.

The linear regression has many advantages - first of all it is fast, both in the estimation of model parameters and its evaluation. Unfortunately, it is also very sensitive to outliers and, of course, it is used to model purely linear (or linearized) relationships. This method should be considered as one of the approaches to the problem of prediction (or forecasting) the solution of which is found as extrapolation of the curve resulting from modeling. In the issue of bots, it is ideal for predicting short- or long-term trends. Due to the rapidity of activity, it can be used, e.g., to determine in real time the increase or decrease in exchange rates, e.g. on the Forex market\textsuperscript{14}, or for predictions based on historical traffic volumes on the Internet or RTB systems (or other simple meter parameters).

The linear regression approach also allows effective aggregation of metrics aimed at measuring the features that allow distinguishing a man from the bot, which are described in Chapter 1. An example of such aggregated measure is viewability which considers both part of the advertising space and the time the advertisement was visible. For specific data and specific applications, it would be possible to, e.g., define a domain/subdomain/host scoring (defining the associated risk of frauds, or the chance that it was created by hackers or spammers) as a linear combination of, among others, such parameters as clickthrough, scrollability and average visit time\textsuperscript{15}.

Also, the problem of identification of bots can be addressed with the use of a certain variant of linear regression, namely the models of the qualitative variable. They are characters models

\[ Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki} + \varepsilon_i, \quad \text{dla } i = 1, 2, \ldots, n, \]

\textsuperscript{13} Wooldridge J. M., Introductory Econometrics: A Modern Approach (Upper Level Economics Titles)


\textsuperscript{15} The given measures should be treated only as an illustration.
where the explained variable $Y_i$ is a qualitative (zero-one) variable\textsuperscript{16}. It is called a **linear probability model**. It is a linear model, the estimation of which uses the Least Squares Method. A matched line is interpreted as the probability of an event that $Y_i$ assumes a value of 1

$$p_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki}.$$  

The adjusted parameters $\alpha_j$ for $j = 0, 1, \ldots, k$ are interpreted as the probability increase associated with the unit increment $X_j$ with the other unchanged values (ceteris paribus). The basic disadvantage of this model is that the values $\hat{Y}_i$ do not always belong to the range $\langle 0; 1 \rangle$, i.e. they cannot always be interpreted as probabilities. In addition, usually the random components are heteroskedastic, while the value of the determination coefficient $R^2$ is small and not very representative. Nevertheless, this model is very fast when it comes to its evaluation, which in the case of the need to solve tasks in real time is of key importance.

The linear probability model can potentially be used to identify bots as a simple, not as accurate but fast classifier. Wanting to get some aggregate (a linear combination of variables) answering the question as to what is the chance that the user with given parameters is a bot. For example, taking into account the variables determining the number of sites visited, the number of activities performed by the user, the average time spent on one site and the longest time from the site loading to clicking, one can choose optimal model parameters that would allow to determine that the bot compared to a human has many times more visited the sites but, on average, it spends much less time on the site than a human being. Using the obtained model as a classifier, we receive a prescription for a quick verification of the “humanity” of the Internet user.

The linear probability model is certainly not an ideal classifier and practically each of the techniques provided in the further part of the chapter is

\textsuperscript{16}In the ordinary linear regression, the variable is usually explained with a continuous, quantitative variable.
characterized by higher effectiveness, but it is a classifier that can provide the first approximation of the solution and then a reference solution.

2.2. Logistic regression

The linear probability model is an example of a qualitative variable model that has many disadvantages. The alternative is a logistic regression model, called in short, a logit model. In the simplest case, this is a two-class classification method that is almost as fast as linear regression. The following relationship is assumed between the probability of an event $P(Y_i = 1)$ and the explanatory variables $X_{ji}$ for $i = 0, 1, \ldots, n$, $j = 0, 1, \ldots, k$:

$$p_i = \frac{e^{\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki}}}{1 + e^{\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki}}} = \frac{\exp(Z_i)}{1 + \exp(Z_i)}.$$

The expression

$$\frac{p_i}{1 - p_i} = \exp(Z_i)$$

is called an odds ratio, while

$$\ln \frac{p_i}{1 - p_i} = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki}$$

is called logit. In the logit model, the parameter sign determines the direction of relationship between the explanatory and explanatory variables. A positive value $\alpha_j$ means that the increase of $X_j$ is connected with the increase of chances that $Y = 1$ (e.g. that we are dealing with a bot). A negative sign means that the increase of $X_j$ is connected with a decrease in $Y = 1$.

The marginal effects in the logit model are not constant and depend on the explaining variables. On the other hand, their interpretation, is similar to the interpretation of parameters in the linear probability model and is read as the probability increment of $Y = 1$ related to the unit increment of $X_j$, ceteris
paribus. Usually, the marginal effects for the mean values of variables $X_j$ are calculated.

In the case of a balanced samples, a value of $Y$ equal to one is predicted if the calculated probability is greater than or equal to 0.5, and to zero value otherwise. For the unbalanced samples, the Cramer rule is often used. If, for example, a sample includes 30% of ones, then we make a forecast of $Y = 1$ when we get a probability greater than or equal to 0.3.

The measure of the accuracy of ex post forecasts is usually a coefficient called count-$R^2$ coefficient, which is expressed by the following formula

$$R^2 = \frac{\text{liczba trafnych prognoz}}{\text{liczba obserwacji}},$$

and it can be read from the contingency table.

The logistic regression model is one of many models of the quality variable, although probably one of the most popular ones. An alternative is, e.g., the probit model according to which it is assumed that the probability values $p_i$ are the values of the $F$ distribution function of the standard normal distribution, i.e.

$$p_i = F(Z_i) = \int_{-\infty}^{Z_i} \varphi(t) dt.$$ 

where

$$Z_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki},$$

and

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

is a function of the standard normal distribution density. The interpretation of parameter signs and marginal effects is identical as in the logit model.
The logit and probit models are usually used as classifiers. They do not have the disadvantages of a linear probability model, and they have many of its advantages. They are commonly used in recommendation systems or credit risk assessment (automatic pre-scoring). However, they can also be successfully used to identify the bots, and more generally to detect any frauds. Those approaches are much less frequently used for prognostic purposes.

There are also other types of regression, except the linear and logistic ones. For example, the Poisson regression is used to detect rare events, i.e. in monitoring or detection tasks. It also seems to have equally great potential in classification tasks as a logistic regression, although due to the popularity of the latter, it is not used for this purpose.

2.3. Decision trees

The decision (classification) trees constitute a completely ideologically and theoretically different approach to the problem of classification\(^\text{17}\). They use the structure of a connected, acyclic graph called a tree. The decision tree consists of nodes, the first of which is called the tree root, and the nodes without further branches are leaves. At each node, a set of features for a given issue is subjected to an attribute value test - it can be, e.g., a comparison to a specific numerical value, or a true-false logical test. Depending on the type of the test, the branches that represent the possible answers to the given question sprout out of the tree root. This process is repeated in each node until the maximum assumed height of the tree is reached, or until the available features are exhausted. In practice, the trees with many levels require a lot of computing power disproportionately to profit from improving the effectiveness of classification. Hence, at the cost of often only slightly worse result in the accuracy test, the height of the tree is trimmed down, saving a huge amount of time and computing power of the computer. This makes avoid the very likely, so-called overlearning of the model.

We can think of the classification process based on the decision tree as of breaking the important decision into a sequence of questions. The presented results are easily interpreted even by people without analytical experience. This is

an undoubted advantage wherever the result obtained by the decision system must be justified by man. The decision trees can be used in almost all classification cases. Therefore, they are an alternative to logistic regression models in the case of the bot identification issues. The predominance of trees is practically a lack of assumptions of applicability, while their disadvantage is a much longer learning time, and above all, a poorer memory and computational complexity when it comes to evaluation. In many cases, heavily unbalanced trees are obtained or their interpretation is far from the intuitive one, due to the discovery of a relationship that is not true, but created only as a response to the specificity of data. The classification trees are often used to assess risk or filter spam. However, they are not suitable for activities in real-time or with limited available resources. Therefore, one can use them to identify bots, however to identify them rather offline. You can expect high effectiveness in such tasks, but possible updates cannot be done too often.

One of the disadvantages of the decision trees is their low stability - even a small change in data can lead to a tree with completely different relationships. The answer to this problem is the modification of the classification tree, which is a boosted decision tree. The tree-boosting technique has evolved as one of the most powerful methods of the predictive data mining. The implementations of those algorithms allow to use them in the regression and classification problems, for quantitative and qualitative predictors.

Within the method of boosted decision tree, a series of “ordinary” trees is created, each of which is built on the basis of the prediction errors generated by the previous ones. Each the subsequent tree is matched to the residuals and determines the next division, which is matched to those residues and determines the next division, at which the variance of residuals (or the error) is even smaller (for a given sequence of trees). Thanks to this approach, one can get a perfect fit of the predicted values to the observed values, even for nonlinear relationships.

---


course, the computational complexity of the boosted decision trees is potentially very large (much larger than the "ordinary" regression trees), because more than one tree is constructed. Therefore, the real-time applicability of that algorithm is very limited.

The random forests are another generalization of the regression trees\textsuperscript{20}. The random forests consist of decision trees in the likeness of nature, where the forest consists of trees. The random forest creation algorithm boils down to the construction of many decision trees, with the number of those trees being defined by the user. A random observation sample (bootstrap mechanism) is selected for each tree, consisting of a fixed number of explanatory variables. The final decision regarding the determination of prediction is made by majority voting conducted on the basis of constructed trees - the more often a given classification result was returned, the greater the chance that it was correct. The main advantage of the random forests is the greater accuracy of the model than in the case of a decision tree. The disadvantage is the loss of the ease of interpretation of the results. We know the result, but we cannot determine what influenced such and not another result.

Each of the methods listed in this subsection: ordinary regression tree, boosted decision tree and random forests, has a similar range of applicability. The rapidity of activity and the simplicity of implementation are in favor of a single regression tree, but the effectiveness is more important in the identification tasks. In the vast majority of cases, both the random forests and boosted decision trees have a higher hit rate than a single, ordinary regression tree. Sometimes the following sequence is given as the real sequence in terms of the prediction accuracy: the most effective are the boosted decision trees, then random forests, and the lowest accuracy of the forecast can be expected from the ordinary regression trees. However, there is no mathematical proof that confirms this rule. It is purely empirical.

2.4. Naive Bayes classifier

The principle of Bayes classifiers uses directly the Bayes theory \(^{21}\) and comes down to predicting categories in an unknown set of data using the probability theory rules.

\[
P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}
\]

\[
P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)
\]

*Figure 3 Bayes formula and principle of operation of the naive Bayes classifier.*


At the same time, a complete independence of explanatory variables is assumed. This assumption is unrealistic and is unlikely to be met. Hence the name of the classifier - naive. Despite this unrealistic assumption, the naive Bayes classifier is often extremely effective in solving classification problems and therefore often used in practice. It has been successfully used in such issues as text classification\(^{22}\), medical diagnoses, and system performance management.\(^{23}\) Irina Rish, in her article entitled “An empirical study of the naive Bayes classifier”\(^{24}\) states that the reason for the success of the naive Bayes classifier is the fact that


\(^{24}\) Rish, Irina. An empirical study of the naive Bayes classifier. Thomas J. Watson Research Center P.O. Box 218 Yorktown Heights, NY 10598. 2001.
the optimality in the case of a zero-one classification does not have to be related to the quality of matching to probabilities distribution. Therefore, the optimal classifier is optimal as long as both the actual and estimated distribution agree with the most probable classes.

This classifier is quite fast in learning and evaluation, which is why it works well on large data sets. Typical applications include: recommendation systems, real-time prediction and, above all, textual analyzes, including spam filtering based on the content of emails, public opinion polls and identification of the Internet trolls by classifying traces of their current activity.

2.5. k-nearest neighbors method

The k-nearest neighbors method is based on the intuitive assumption that neighboring objects are similar to each other. It is used in the forecasting statistics, but it can also be used for classification. The next steps of the algorithm for a given x observation are as follows:

1. comparing the values of explanatory variables for a given observation x with the values of those variables for each observation in the training set.
2. selection of k- observations from the training set nearest to x.
3. average value of the explained variable for the selected observations.

We assume the value obtained in step three as the forecast. However, the result depends on the measure of distance adopted in the second point. The most commonly used metrics are the Euclidean, Manhattan and Chebyshev ones. It is also a non-trivial task to find the optimal value of parameter k.

For the issues described in this study, the k-nearest neighbors method would be an ideal tool to categorize the users by their behavior or domains by their statistics. The disadvantage is the relatively high complexity of the algorithm which increases rapidly with the increase in the number of clusters into which we want to divide the examined set, with the increase in the number of dimensions (factors studied), as well as the number of observations. A characteristic feature is

25 https://www.statsoft.pl (last accessed date: 06/01/2019).
the nondeterminism embedded in the starting values of the method. They are usually random. This means that one can get different solutions for different draws. It depends on the specific case whether this is an advantage or a disadvantage. However, the method itself certainly allows for a multidimensional analysis of the problem under consideration. In each case, the results are interpretable, but with a larger number of dimensions, the visualization of the solution is difficult.

In the case of bots, the k-nearest neighbors method to divide the set into clusters of elements from that set can be used in many different ways. The simplest way is to divide the set into two subclusters: one interpreted as bots, the other as people. The statistics described in Chapter 1 can be selected as factors. In contrast to the logistic or linear regression, the factors are not aggregated into one number, corresponding to the scoring of a given observation. Here the points are not subjected to any transformations. A plane is found that makes it possible to separate those two clusters of points from each other.

The results of the comparison of the effectiveness of selected machine learning algorithms in the botnet detection task based on the empirical analysis presented by Melisha Dsouza are included in Table 2. The data from the table shows that many of the mentioned methods can achieve unusually high effectiveness (over 95%) in terms of the detection issues. In the other listed issues, the results would most likely be different. However, this is, of course, an off-line analysis of the already cleaned data set. With the real-time detection issues, one often has to compromise and at the expense of effectiveness, improve the performance of the algorithm used.

---

Table 2. Results of the botnet empirical detection.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>99%</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>96%</td>
</tr>
<tr>
<td>Naive Bayes classifier</td>
<td>72%</td>
</tr>
<tr>
<td>K-nearest neighbors method</td>
<td>100%</td>
</tr>
</tbody>
</table>

Source: own study.

2.6. Artificial neural networks

The artificial neural networks\(^{27}\) are modeled based on the human brain and constitute one of the basic algorithms of the machine learning. They are not a new idea because they were invented over 70 years ago. However, to this day they are one of the most sophisticated algorithms used in both the classification and regression problems. In recent years, their popularity has increased with the increase in the popularity of deep learning.

The diagram of the neural network (see Figure 4) is described by means of an acyclic directed graph. The main element of the neural network is the processing neuron. There are many neurons in the network that have any number of inputs and outputs. Neurons are clustered into layers in which each neuron is connected to each neuron of the preceding layer. Thus, the values of variables are transmitted in a progressive way between individual layers of the neural network. In the next layers, operations on variables are performed until the result value is reached at the end of the graph.

\(^{27}\) Haykin S. Neural Networks and Learning Machines, Prentice Hall, 2008.
The artificial neural networks are applicable, among others, for problems in which one cannot recognize any (simple) pattern. Their drawback is a very long calculation time, hence on the one hand a smaller number of observations speeds up the learning time, and on the other hand, a large number of observations improves its accuracy. Neural networks, apart from those that have fewer layers and fewer neurons in the layer, are hard to use in tasks whose solution must be found in real time. The performance issues are a problem and they are similar to those that are encountered in the decision trees. The result returned by the artificial neural network is generally difficult to interpret - this is another drawback of the artificial neural networks. The artificial neural networks, due to their structure and application of the learning methods, can be used wherever certain known information is inferred from certain unknown information. The advantage of the artificial neural networks is the ability to find an approximation of the solution even in situations where no relationships or patterns are known. The network will not bring us closer to knowledge about them (see the difficulties with the interpretation of network results), but we will get an effective machine to receive them.

Typical issues in which artificial neural network algorithms are used include the prediction issues such as stock price or stock market indices forecasting, as well as the classification issues, such as borrower scoring, spam filtering, user categorization. An example of empirical analysis of troll detection on Twitter.
showed that the artificial neural network achieves the effectiveness at the level of 95%. Certainly, the scope of applicability of the artificial neural networks will grow with the development of tools for effective processing and storage of large amounts of data.

2.7. Panel models

Thanks to modern data processing techniques, the opportunities to improve the effectiveness of econometric modeling appeared. In many cases, we have panel data, i.e., the data that has the features of cross-sectional data (describing the cluster at a single moment) and features of time series (describing the unit in different periods). The potential use of such data should provide the estimates more precise than in the case when one of the dimensions was omitted. We could then lose information about the differentiation between individual units, or data on variability over time. This approach is successfully used, e.g., in social studies, medicine, but also in the analysis of the behavior of customers of insurance companies and banks. However, until now it has rarely been used in the strictly IT issues. Presumably, the panel methods will also enter the world of web analysis, but so far there are no tools that allow easy operation on large data sets. A few implementations of the panel estimates are available. One of them is the fixed effects estimator using the MapReduce programming model, implemented in Apache Spark.

An interesting example of the use of the panel techniques is, e.g., the method of detecting fraud attempts in the Real Time Bidding system proposed by Bernardelli (2015). The method consists of two different models, one designed for

---

28 https://medium.com/@conspirator0/identifying-political-bot-troll-social-media-activity-using-machine-learning-20dcd56e961a (last accessed date: 06/01/2019).


the classification of users, and the other one for distinguishing real websites from those specially planted by fraudsters. The presented models are closely related and together they seem to be quite a fast and effective tool to separate the human Internet traffic from that generated by bots. In an ordinary, non-panel approach, the web sites (the planted ones vs. real ones) and the users (people vs. bots) would be modeled separately. Of course, one could also achieve high classification effectiveness in those cases, but using information about sites that are visited by the bots and users who visit the planted websites, one can quickly detect a new fraud by observing the sites or users that he already knows that they are an attempt to cheat. The effectiveness of the proposed method has been verified by computer simulations.

2.8. Machine learning

This sub-chapter is devoted to the presentation of methods useful in the identification of Internet bots concerns not as much the technique as such, but the whole field of the machine learning which is part of the artificial intelligence problem. The following is a selection of applications (according to Wikipedia) of the machine learning algorithms in the category of data analysis and classification:

- systematics of astronomical objects (NASA Sky Survey),
- diagnosis of diseases based on symptoms,
- modeling and development of drug therapies,
- handwriting recognition on the basis of examples,
- classifying data into thematic clusters according to criteria,
- approximation of an unknown function based on samples,
- determining functional relationships in data,

---

• predicting trends in financial markets based on micro and macro-economic data,
• detection of money laundering.

It is worth noting that the machine learning is constantly evolving and finding new practical applications. The potential of the machine learning methods is enormous. You can come across - still relatively few - implementations using tools of this field of science. Many of them involve attacks from the botnet network - their detection and analysis of operation.

A large, still unused part of the works that could significantly improve the effectiveness of the existing algorithms for monitoring, classifying and filtering various parts of the Internet, is associated with advanced lexicographic analysis and automatic processing of natural language. It is easy to imagine that such an analysis of texts on websites created for scams or spam messages should give a sensible contribution to knowledge about patterns or features that allow them to be easily identified. An interesting example in which a modern analytical approach is used is the method of identifying the gender of the Internet users using data from the users profiles containing the website addresses and the frequency of visits. The proposed approach combines the difficulty of Big Data processing, lexical analysis of words from the Internet domains (the Word2Vec algorithm), artificial neural networks, a mathematically sophisticated vector representation of the user profiles and logistic regression as the main classifier. An empirical analysis was also performed, achieving the effectiveness of classification at the level of 82%. The method itself can be used, e.g., in personalized marketing as a source of savings in the form of reducing unnecessary expenses for mistargeted advertising.


37 https://radimrehurek.com/gensim/models/word2vec.html (last accessed date: 06/01/2019).
2.9. Unsupervised learning

It is assumed, based on a number of empirical studies, that the supervised learning has greater effectiveness than the unsupervised learning. However, in the case of the detection of bots, the setting of some parameters, without access to the test set, is at least strongly hampered. Therefore, it seems crucial to develop such algorithms the effectiveness of which is not based on prior learning on the training sets. Below are two suggestions, using the algorithms provided in the previous sub-chapters.

The first of the suggestion is based on the analysis of clusters in relation to the measures given in the previous chapters, e.g. number of activities, number of views, number of posts, number of followers, times of views or execution of activities, viewability, etc. This analysis would be a multidimensional analysis and would allow effective creation of sensible clusters related to the features of activities in the Internet. The use of an appropriate clustering algorithm seems to be key here and the efficiencies achieved can vary significantly. Therefore, it is recommended to check a number of classification algorithms and perform a certain averaging or select the most appropriate one (depending on the results obtained). The proposed specialized algorithms (apart from the ordinary k-nearest neighbors method) include clustering algorithms used in Internet search engines:

- Agglomerative Hierarchical Clustering (AHC);
- Lingo;
- Rocchio algorithm;
- TC algorithm;
- STC algorithm;
- LSA;
- PLSA;
- WebSOM;
- QDPageRank;
- SVD.
Pursuant to a number of articles\(^ {38}\), the predicted effectiveness of such methods oscillates between 80-90% if the data covers a broad cross-section of the Internet traffic. On the other hand, it is difficult to predict the effectiveness when the data stream is limited to several publishers. On the one hand, general algorithms should have much lower effectiveness (approx. 5-15%), on the other hand, there is the possibility of creating dedicated classification methods per publisher, which could ultimately be more effective than the data-based approaches.

The second of the proposed methods is based on the text mining. According to common sense, the information placed by bots should contain words that are over-represented in relation to messages written by people. For example\(^ {39}\), comparison of the frequency of words in favorites on Twitter for bots and real people is presented in Figures 5-6. Certainly, differences should be noticeable, although it is necessary to track changes over time. The problem in assessing the effectiveness of text mining methods is the use of Polish, which is syntactically more difficult than English. Thus, the potential of this approach is visible, especially if the context analysis and the latest implementation algorithms are also implemented (see, e.g., word2vec, Levenshtein distance, convolutional neural networks).

It is also worth considering a combination of text examination and certain measures measuring approaches. An interesting approach to the detection of bots on Twitter was applied in 2018 Efthimion\(^ {40}\) et al. The variables proposed by them that can be used in the bot detection procedures are listed in Table 3. Using the support vector machining approach, they achieved over 95% effectiveness (on selected datasets, supervised learning, English). Of course, it does not seem possible to achieve such effectiveness in the overall case, but it presents the potential of this approach.


\(^{39}\) Cohen D., Text Mining: Twitter and the Problem With Spam Bots.

Figure 5. Frequencies of words in Favorites (Twitter) for bots.

Source: Cohen D., Text Mining: Twitter and the Problem with Spam Bots.

Figure 6. Frequencies of words in Favorites (Twitter) for people.

Source: Cohen D., Text Mining: Twitter and the Problem with Spam Bots.
Table 3. A proposal to use variables to classify bots.

<table>
<thead>
<tr>
<th>Area of Analysis</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile</td>
<td>Absence of id</td>
</tr>
<tr>
<td></td>
<td>Absence of a profile picture</td>
</tr>
<tr>
<td></td>
<td>Absence of a screen name</td>
</tr>
<tr>
<td></td>
<td>Has less than 30 followers</td>
</tr>
<tr>
<td></td>
<td>Not geo-located</td>
</tr>
<tr>
<td></td>
<td>Language not set to English</td>
</tr>
<tr>
<td></td>
<td>Description contains a link</td>
</tr>
<tr>
<td></td>
<td>Has sent less than 50 tweets</td>
</tr>
<tr>
<td></td>
<td>2:1 friends/followers ratio</td>
</tr>
<tr>
<td></td>
<td>Has over 1,000 followers</td>
</tr>
<tr>
<td></td>
<td>Has the default profile image</td>
</tr>
<tr>
<td></td>
<td>Has never tweeted</td>
</tr>
<tr>
<td></td>
<td>50:1 friends/followers ratio</td>
</tr>
<tr>
<td></td>
<td>100:1 friends/followers ratio</td>
</tr>
<tr>
<td></td>
<td>Absence of a description</td>
</tr>
<tr>
<td>Text Analysis</td>
<td>Levenshtein distance between user’s tweets is less than 30</td>
</tr>
</tbody>
</table>


Table 4. A proposal to use variables to classify bots.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Misclassification Rate</th>
<th>True Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Spambot</td>
<td>95.77%</td>
<td>4.23%</td>
<td>96.81%</td>
</tr>
<tr>
<td>Traditional Spambot</td>
<td>96.25%</td>
<td>3.75%</td>
<td>97.13%</td>
</tr>
<tr>
<td>Fake Followers</td>
<td>100%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>NBC News Russian Bots</td>
<td>99.87%</td>
<td>0.13%</td>
<td>98.91%</td>
</tr>
<tr>
<td>Total</td>
<td>97.75%</td>
<td>2.25%</td>
<td>98.98%</td>
</tr>
</tbody>
</table>

Imprint

Author:  Michal Bernardelli
Published by:  ABT Shield
             Edge NPD Sp. z o.o.
             22A Czeska Street
             03-902 Warsaw, Poland
             office@abtshield.com
             http://abtshield.com/