Zeszyty Naukowe Wydziału Informatycznych Technik Zarządzania
Wyższej Szkoły Informatyki Stosowanej i Zarządzania
„Współczesne Problemy Zarządzania”
Nr 1/2021

DISTINGUISHING THE ARTIFICIAL AND THE GENUINE AD-RELATED TRAFFIC: MAIN OBSERVATIONS AND EXEMPLARY RESULTS[1]

Marek Gajewski[1], Olgierd Hryniewicz[1], Agnieszka Jastrzębska[1], Mariusz Kozakiewicz[2], Karol Opara[1], Jan W. Owsiński[1], Sławomir Zadrożny[1] and Tomasz Zwierzchowski[2]

[1] Systems Research Institute, Polish Academy of Sciences;
[2] EDGE NPD Ltd. Co.
Warsaw, Poland

Abstract: We describe the essential aspects of the project, aimed at developing a methodology for distinguishing the artificial, i.e. automatically generated, internet traffic, from the genuine ones, i.e. produced by humans, regarding the advertising on the web. So, we first present the nature of the problem, including its rough business rationality and then the key characteristics of the relevant internet traffic. This is followed by the outline of the set of methodologies used on the project, and then an excerpt of the results is presented with the corresponding comments and discussion. Finally, the conclusions, technical and of a more general character, are forwarded.

Keywords: internet, advertising, artificial traffic, bots, classification, clustering, data analysis.

1. Problem outline

1.1. The advertising market on the web

Advertising constitutes, undoubtedly, one of the primary moving forces of the world-wide-web, or the internet. One can hardly imagine what share of the web content would remain were the advertising removed or banned, but, definitely, we would have been dealing, then, with just a shade of what we have today (“money

[1] The work reported here was carried out in the framework of the project named *ABTShield*, implemented by EDGE NPD Ltd. Co. in cooperation with the Systems Research Institute of the Polish Academy of Sciences under the regional operational program of the province of Masovia in Poland, financed from the European Regional Development Fund, contract no. RPMA.01.02.00-14-B448/18-00.

makes the world go round…"), while, definitely, fake news, stupidity and aggression would not go away with the demise of advertising, given the facility of "expression", provided by the web, the emotions of social, political, racial etc. origins, and the wish to exploit all these.

The advertising activity (which, in fact, is also largely associated with marketing, since it is not just a passive provision of content, as this takes place, e.g., within the outdoor advertising activity), taking place on the web, is performed in the framework of an extensive and quite complex market. The organisation of this market involves several distinct roles. Although the different actors, functioning on this market, can take on more than just one role, the essential roles can be roughly outlined as follows:

- The advertiser: the entity that wishes to have its product, service, brand, image etc. promoted, made visible, offered for sale, …, and that notwithstanding whether the actually intervening entity, taking part in the process, is a true producer, or distributor, or sales agency.
- The advertising agency: this body may be responsible for a number of activities, such as advertising design, campaign design, campaign organisation, monitoring and verification etc.
- The demand side platform: these platforms, which interact with the subsequent kind of actors, make it possible to technically organise the flow of advertising content for placement in the web media.
- The supply side platform: this kind of actor is responsible for actual placing of the content with definite kind of medium. The placing is the effect of functioning of instruments, largely based on real-time auctions[2], putting together the demand and supply sides along with pricing and other characteristics of the displayed advertising material.
- This advertising material is then displayed by the media, run by their owners or operators, the material being channelled through respective advertising tools of a given medium, platform, etc.
- Finally, this material is seen, displayed, clicked on, etc., by the ultimate "customer", this notion being insofar misleading as we deal, in fact, in this particular study at least, only with a potential customer, whose behaviour, not necessarily involving purchase, is the ultimate objective and the yardstick of the whole system.

Thus, the system is quite complex – even if apparently "linear", with the "line" leading from the advertiser to the customer. Actually, it can, and preferably does, turn into a circular one, with customers, at the far end, effectively purchasing the products

[2] The web users are, as a rule, not aware of the fact that while they move to a given web page, which is supposed to provide the advertising content, their properties (as expressed through, in particular, the "cookies") guide the flash auction, resulting in the advertising material they will see appearing on the monitor screen.

or services, or at least actively inquiring about them (“conversion”). The complexity of the system results from, first, the multiplicity and the diversity of the actors, playing the particular roles, the diversity of content, the dynamic character, and, finally, the possibility of introducing the deformations to the functioning of the system, and the market[3].

With respect to the latter, the present paper reports on the work done within a project, which addressed exactly the primary manner, in which the web-based advertising market can be twisted. Namely, it is usual for an ad product to be paid for in terms of a certain (minimum) number of clicks on it (“how many times it has been effectively shown”). Hence, if an automatically functioning software robot (“bot”) clicks a certain number or proportion of times on the ad, objectives are achieved of definite market value for several (kinds of) participants of the market game. It is known that such a traffic exists and there are (quite extreme, in fact, and certainly worth verifying) opinions that it may even reach around 50% of total traffic volume (see, for instance, the website https://www.cheq.ai/blog/what-is-bot-traffic).[4]

1.2. Artificial activity and its general characteristics

The artificial activity as mentioned before pays in some way, either positive (this being the reason for its development) or negative, regarding individual participants of the web-based ad market. Thus, while it is a loss to the ultimate advertising agent (producer or seller of goods, or both), and perhaps also to the designing agency (less of the actual views by humans), it is a bonus for those, who get paid for a definite number of clicks (various intermediaries), and, in a specific manner, to the competitors of the advertiser. Beyond this, a secondary (or “underpinning”) market would develop of instruments and tools for operating on the primary market, i.e., the one for producing the respective bots or robots, and also for counteracting, in some manner, their activity. One of the components of the latter is the *capacity of estimating the dimensions of the artificial traffic, or even better: identifying the artificially produced events*, this capacity having also a definite market value (e.g., reliability of invoicing for the ad campaigns), depending upon its accuracy and verifiability.

In a different paper, reporting on some aspects of the here considered project (Gajewski et al., 2021)[5], a bit more space is devoted to the characteristics of the market, especially in terms of measurement of effectiveness of advertising, which, obviously, depends on the cost and price levels, on conversion rates, i.e., progress from clicking on the ad towards the actual purchase, and, exactly, on the share of the

[3] It can be easily deduced that in case of observing the actual conversion (purchase) the room for manipulation becomes very narrow, indeed.

[4] Note that we do not consider here the crawlers and bots that have no “negative” objectives, like those that gather statistical data, are used for scientific purposes etc.

[5] See also Gajewski et al. (2022) for an already published short account on some of the essential characteristics of the study here described.

artificial traffic, which, definitely, remains an unknown (conversion rates being estimated in a rather straightforward manner, even if also with some error). Some attention was also devoted in that paper to other phenomena than the one considered here (bot clicking), standing in the way to a proficient operation of the web ad market.

We shall not go deeper into the working of the market in question, although, definitely, if some at least of its parameters (like all "prices") were known, one could try to simulate, if not solve analytically, the respective game and try to find the reasonable solutions, or at least the conditions for such solutions, indicating the potential behaviour patterns of the participants. Yet, it is obvious that some of the (interrelated) basic characteristics of the corresponding artificial traffic, related to the obvious market prerequisites (profitability), are:

- Low cost of individual action, and
- Low cost of producing and operating the instruments, with
- Large scale (or numbers) of interventions.

In the above context the fundamental question arises of the difficulty of identifying the artificial traffic. If this were really simple (and hence cheap), the bots would have been easily washed away from the market, which would have then be a "clean" market, i.e. the end customers would then really pay for true glimpses of the ads they ordered. However, there are two reasons that make the identification of artificial traffic somewhat more complex: **1**. The non-deterministic character and the very high variety of human behaviour events and patterns; coupled with **2**. The relatively low cost of producing bots that do not act truly "weird".

Yet, notwithstanding the truth of the latter statement, the operation of such automatic instruments must remain relatively simple (but, definitely, not "simplistic"), because: **a**. production of an instrument that would really try to imitate human behaviour (and then its more complicated operation) might be more expensive, and **b**. such an instrument would presumably not produce the required scale of interventions (just because it would behave like a human).[6]

Altogether, against the background of these remarks, one can easily imagine at least some characteristics that ought to allow for the distinction of the two kinds of traffic. We shall devote to them the subsequent section of the paper.

2. Some key characteristics

The objective of our work was to devise a mechanism that differentiates humans from bots based on website activity logs. The method was delivered to be used by publishers who alongside their standard content display advertisements and

[6] We do not enter really into this discussion here, since it would have to encompass a number of questions, to be potentially answered through appropriate studies, like: what if we simulated a very big mass of humans (cost vs. statistical features)? or: what does it mean to "behave like a human" – does some kind of lack of regularity suffice?

wish to analyse the traffic on their end. Tracking the nature of actors (bots or humans) that are displaying advertisements is a cornerstone for fair settlements in pay-per-click billing schema. There is an organic, and quite understandable in view of the market functioning, movement in the community of stakeholders of the digital advertisement market that is concerned with the credibility of the publishers involved and advocates for more transparent, but also possibly reliable information about the share of bot clicks on their advertisements (Mouawi et al., 2019). The proposed solution can be used for this purpose providing that an independent party will have access to raw data (logs) necessary to analyse the share of human and bot activities on a given publisher's website. In addition, it can be used by publishers autonomously, to gain insight into the nature of the traffic on their pages.

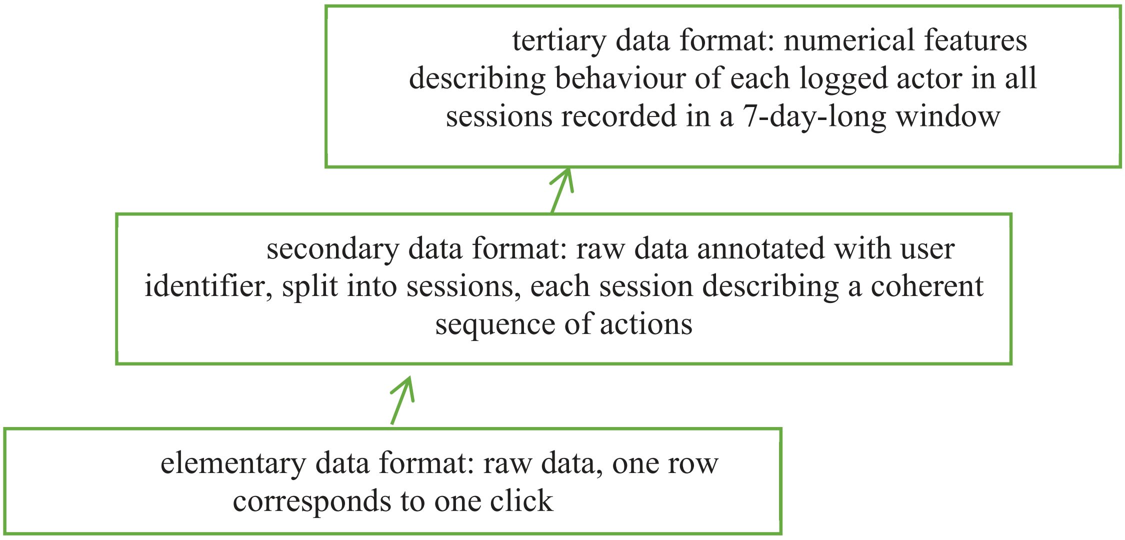
The difficulty of the task of bot detection is increased by the so-called botnets, in which a click fraud is organised in a distributed manner (Khattak et al., 2014), and bot farms, where humans are hired to click on advertisements (Thejas et al., 2021).

The task would have been, of course, much easier, were ampler information available on behaviour of the given "agent", be it just what is s/he doing on the web. The basic question is: what is this agent doing besides clicking on these ads on given websites? Having appropriate answer to this question would bring the error in distinguishing humans from bots very close to zero. Yet, not only would this mean actual tracking of the agents, including humans, and thus actual infringement on privacy, but also two essential issues will have to be rationally resolved: (1) how far should we go in identifying behaviour patterns (how long sequences of events and of what character would have to be available) in order to get "sufficient information"? and, very closely linked: (2) would the results really justify the additional cost of data acquisition and analysis (i.e. would the improvement of the distinction sought over methods based on less information pay for the extra cost)?

In our other paper, already referred to, Gajewski et al. (2021), we also made some remarks on the ways to struggle with the bot activity and to neutralise the effects of this activity, in virtually all cases necessitating a deeper knowledge and capacity of action than we envisage here.

Thus, the proposed solution, or solutions, described here, analyse primarily the temporal characteristics of activities performed by actors accessing a given webpage. As of the time of this writing, we analyse data using the 7-day-long time window. The proposed approach starts to process information at the elementary level, that is – raw data organized into rows, where one row corresponds to one event registered in logs. The most fundamental functionality of the developed approach is an external, proprietary algorithm that is used to assign each event to one actor. Each actor accessing a given webpage receives a unique identifier. This task, however easy it may sound, is in reality quite a challenge. The algorithm must include reconnecting actors with dynamic IP addresses and changing details of their user-agent (technical environment). Subsequently, we produce an intermediate, secondary, data format, where raw data is annotated with user identifiers and split into chunks called sessions. Each session describes a coherent sequence of browsing activities for a given actor, sorted according to timestamps. Finally, we proceed to the extraction of numerical features describing the behaviour of each actor in the registered sessions. In this last data representation format, one row corresponds to one actor distinguished by its unique identifier. Altogether, in effect, we have as many rows as distinct actors who accessed the webpage in question during the analysed 7-day-long window. The described interchanges between information representation formats are illustrated in Figure 1.

Figure 1. Subsequent data transformations[7]


tertiary data format: numerical features describing behaviour of each logged actor in all sessions recorded in a 7-day-long window
secondary data format: raw data annotated with user identifier, split into sessions, each session describing a coherent sequence of actions
elementary data format: raw data, one row corresponds to one click

The discrimination between humans and bots is performed with the use of the final data representation format, which contains numerical descriptors of the actor's behaviour. The features were hand-crafted with the use of expert knowledge of this domain. The main descriptors that can allow distinguishing human and artificial traffic are utilizing the following notions:

- Quantity of clicks;
- The regularity of clicks;
- Frequency of clicks (we distinguish the two in view of the different time frames that may intervene, see also below);
- Unusual activity hours, computed with the use of knowledge concerning the time zone of the actor (e.g., nighttime visitors);
- "Logic" of clicks – i.e. is an actor using referrals when clicking on the next items;
- Changes of user agent details and the agent details themselves;
- Changes, concerning behaviour characteristics, especially the deep and significant changes;

[7] All illustrations in the paper come from own research of the authors

- Frequency of cookie rotation that may indicate suspiciously high browsing activity.

The here presented notions, as can be easily seen, are to a large extent simply a common sense choice, given the limited nature of information available. They rely on information about the agent's time zone, technical properties of the machine, from which the agent has established a connection, but to the greatest extent, we utilize information about clicks, as this is the activity that we need to trace in the first place. The selected features account for various *timescales*. The respective values are computed using data concerning sessions that were registered in the time frame of 7 days. The features describe the behaviour of an actor within sessions, but also relations between consecutive sessions, the rationale being as follows: the simplest bots may be programmed to perform quick actions within a single session, while more sophisticated bots may be programmed to perform actions at certain intervals, which will get recorded in several sessions. The total number of clicks in the latter scenario is still vast, but it is split to conceal the true nature of an actor.

The features actually considered should, according to our expectations, allow for detecting the behavioural patterns of actors accessing the webpage. A similar focus is mentioned in several other studies on bot detection, for example those by Aberathne and Walgampaya (2018) or Cai et al. (2020).

In the proposed approach, we use the set of the following groups of features:

- **Concerning the technical environment of the actor**: distinct user agent count, is the actor's declared browser name the same as the true browser name, is the actor's declared browser language name the same as the true browser language, is the actor's declared operating system name the same as the true operating system name, is the actor's declared screen resolution name the same as the true screen resolution, distinct IP address count, average time in seconds between events happening within a cookie in the last 7 days, variance (in seconds) of the time lapse between changes within a cookie, the number of assigned cookies for the past 7 days, whether the actor is on a whitelist of useful bots (helpers in indexing), the ratio of the number of events logged with whitelisted IP to the number of events logged with a not-whitelisted IP.
- **Concerning activities within a single session**: average time in seconds to the next event in the sessions with more than three events, the maximum number of distinct web pages visited within one hour, the same but within a one-minute window, the maximum number of page views in a one-hour window, the same but within a one-minute window, total page views count.
- **Concerning relations between sessions**: the number of sessions with more than three events, the average time in seconds between consecutive events occurring in the last 7 days, variance (in seconds) computed for the number of events is sessions longer than three events, variance (also in seconds) of the period occurring between events in sessions longer than three events.

- **Concerning the nature of clicking**: distinct page views count, referrals share in all clicked URLs, referrer count, the number of visited campaigns, the number of unique renderers.
- **Concerning actor's time zone**: the number of page views in hours 0-6 in the actor's time zone (hours 0-6 are supposed to be of rather low activity for humans), the share of actor's page views in hours 0-6 per the total number of page views.

The proposed set of features emerged from a series of experiments and analyses on several data sets. The initial set comprised well over 100 features (actually, at some point, even more than 400) and was gradually reduced to the set shortly characterised above. Still, the given set of features shall be critically analysed and there is a potential for its further reduction, depending on the properties of a particular data set one needs to process (see further on this).

Each row in the final data frame is labelled as either a bot or a human. The label is attached using a rule-based method developed by the experts. The rule comprises several compound if-then clauses. The rule is general enough to be applied to different data sets, but it is not an oracle that never misses. We use those labels as an approximate reference to guide our data-driven approach to the mining of this data set.

Let us repeat at this point that the set of variables we use, characterising the web-related behaviour of particular agents, is highly limited in technical terms, and this quite on purpose, as explained. In fact, we refer to very few such basic behaviour characteristics, treating as relevant features their various kinds of aggregates and (their) mutual relations. The consequence thereof is, naturally, a high risk of having very high levels of correlation between the features and low effectiveness of individual features, meaning, actually, that, on the one hand, we could use just a couple of these features without any, or with very little loss of effectiveness, and, on the other hand – this effectiveness, without deeper knowledge, which we assume is not available to us, is quite low. An important part of the respective research was exactly devoted to verifying whether an ingenious use of the features designed and selected can neutralise this risk.

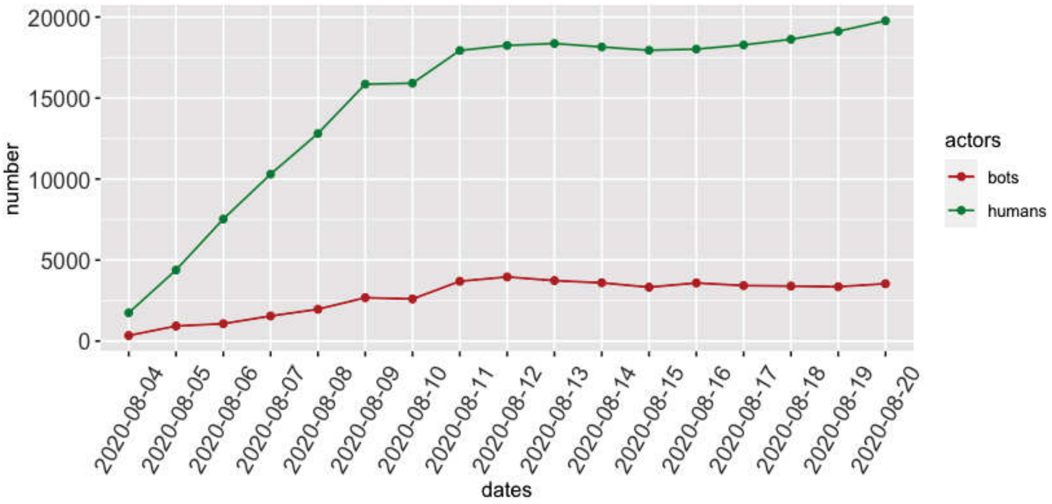
3. Feature analysis - a case study

3.1. Some temporal characteristics

To start with the presentation of the work done, let us introduce a brief case study, in which we will showcase the previously characterised set of features on a real-world example of a marketing campaign promoting one specific product. For the sake of anonymity, we do not give further details on the matter of the company and the product being advertised. The campaign was delivered in the form of banners. We trace the campaign in 7-day-long windows starting from the beginning of this campaign (4th day of August). Thus, we have the opportunity to examine aspects such as the popularity of the campaign.

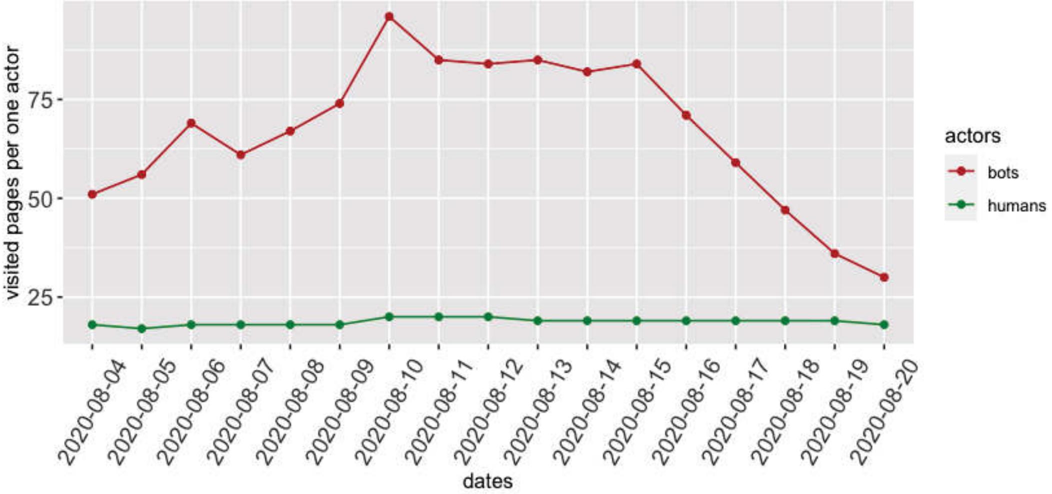
In Figure 2., we can see how many bots (red) and humans (green) were registered in subsequent 7-day-long windows starting from the beginning of the campaign. The respective labels, as mentioned, were attached using our expert-made rule.

Figure 2. The numbers of bots (red) and humans (green), registered in the subsequent 7-day-long windows starting from the beginning of the campaign



So, in Figure 2. we see a gradually increasing interest in the campaign, both on the side of humans and bots. The red line, corresponding to bots seems to be flatter, but it is due to a rather large scale of the OY axis. In fact, both in the case of humans and bots, we see a linear increase and then a stabilization around August 11th. We see that there is an overwhelming majority of humans accessing our site. However, let us look at the key specifics of bot and human activities – the number of page views. We illustrate this in Figure 3.

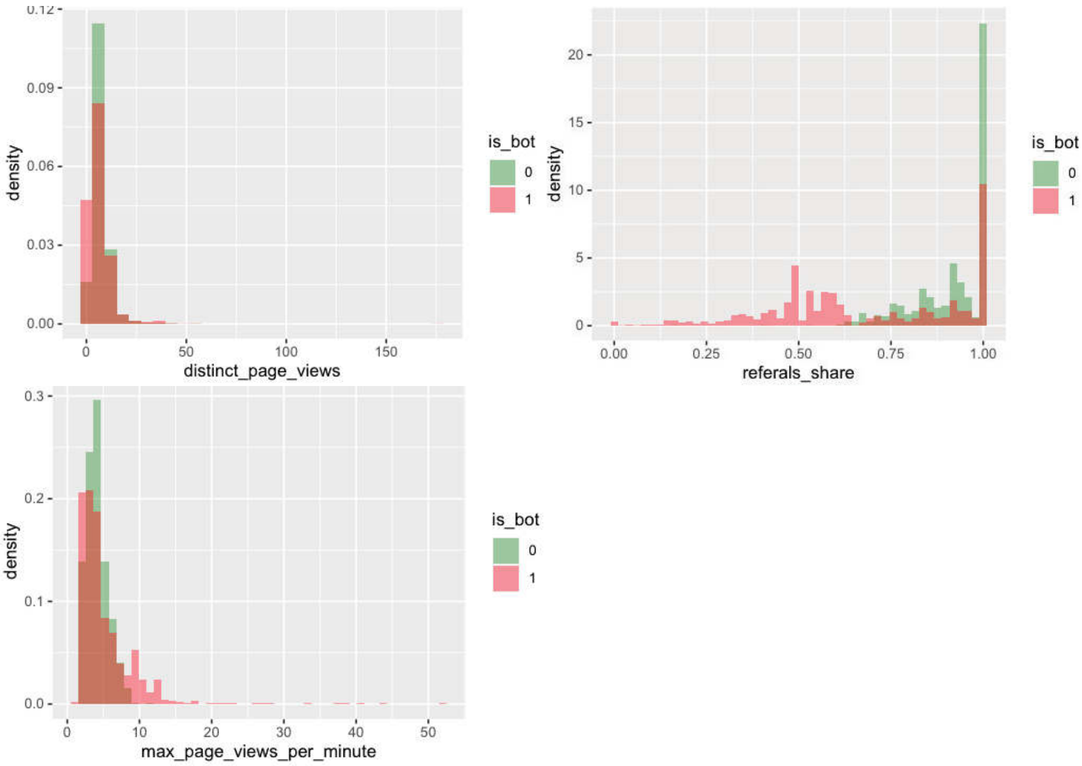
Figure 3. The number of displayed pages by bots and humans divided by the number of, respectively, bots and humans.



Then, in Figure 3., we see that bots are on average substantially more active than humans. Moreover, there are changes in the tendencies when we compare results for different time windows. All in all, bot traffic is dominating the human traffic in terms of intensity.

It is not only the number of page views that seems to allow for distinguishing bots from humans. Thus, in Figure 4., the histograms of values of selected features for the 7-day-long window starting on August 11th are shown.

Figure 4. Histograms of selected features concerning bots (red) and humans (green) logged during a 7-day-long window starting from August 11th.



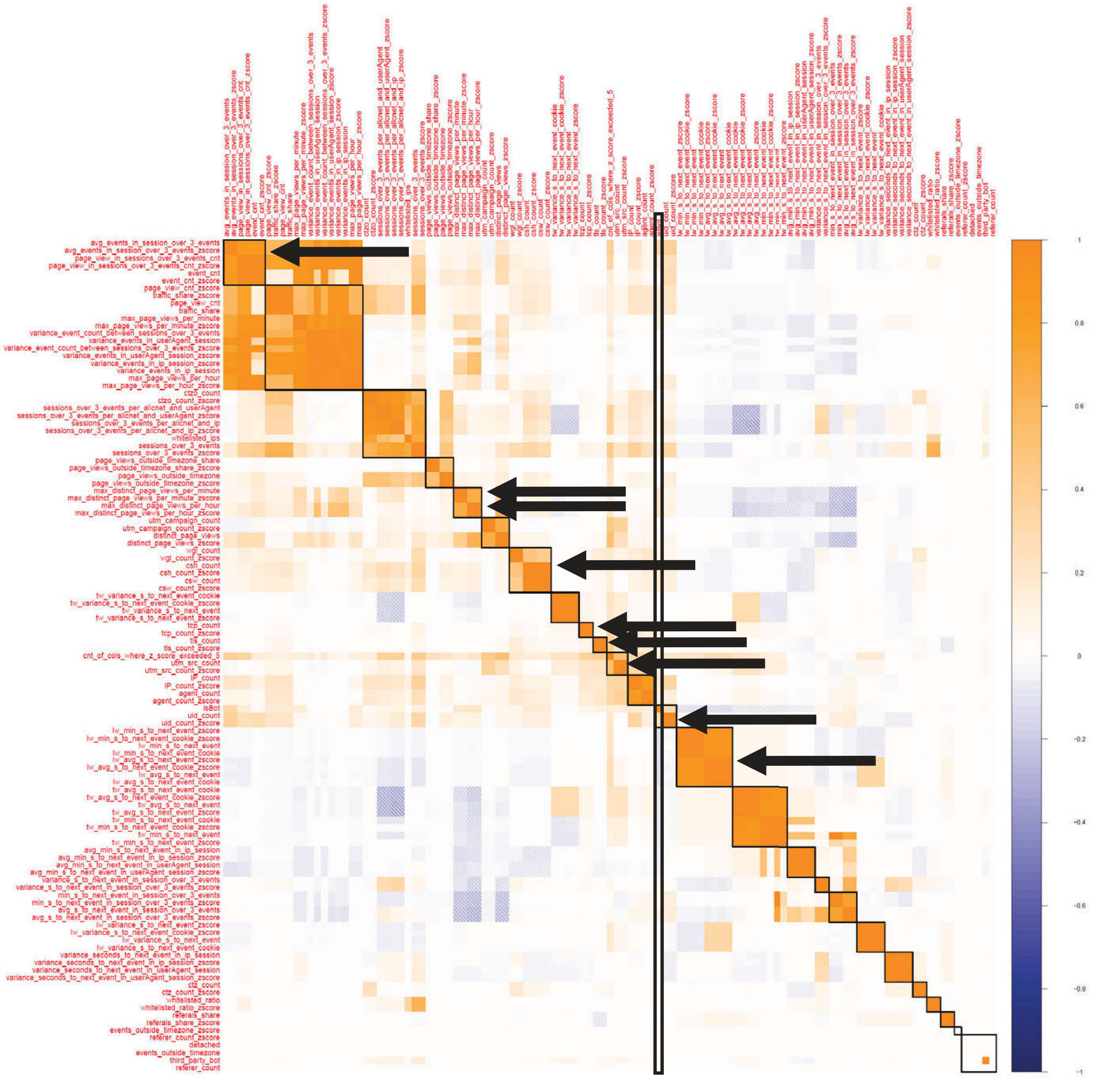
The task of distinguishing humans from bots is challenging. Histograms use slightly transparent colours for the reader to see overlapping values. This shows that there is no clear split between the two types of actors, at least not in terms of the here considered characteristics. Nonetheless, we might try to formulate certain (even if weak) observations. So, we can see in Figure 4. that in the case of some features, like max_page_views_per_minute the distribution concerning bots is slightly more skewed to the left what suggests that the majority of recognized bots view more pages than humans. The same properties, only slightly less visible concern the distinct_page_views feature. In other cases, like in the case of referals_share feature, the distribution of bots is substantially different from the distribution of humans. Thus, one can identify definite differences as to the distribution of the two kinds of agents, but the conclusions are not straightforward in any way.

3.2. Correlation analysis

We shall now turn to the analysis of correlations between the variables considered in the study. Several analyses were performed, mainly oriented at the possibility of well-founded selection of variables to be used in further study. It was, of course, expected that in view of the fact that only a very limited number of original characteristics was used, forming the features considered through various aggregations and mutual relations between them, there should be a high share of high correlation coefficients, in absolute terms.

One of the starting points for the correlation analysis was constituted by the set of 110 variables, used in the expert-developed tool for labelling humans and bots. The result of analysis is shown in Figure 5. below.

Figure 5. Correlation analysis for 110 variables used to tell bots from humans



The image of Figure 5. shows the higher correlations where the colours get more intensive (beige for positive, blue for negative correlations). A very clear block-diagonal structure could be created by the appropriate permutation of rows and columns. The long vertical rectangle shows the labelling variable is_bot. The CFS method (correlation feature selection) was applied to these results and the outcome, i.e. the variables selected by the method are indicated with big arrows. The general principle is that the selected variables ought to be possibly strongly correlated with the labelling variable, but possibly weakly correlated among themselves.

Table 1. shows the variables from Figure 5., featuring the strongest correlations with the labelling variable. Those with correlation coefficient values in darker green colour were selected by the CFS method. Even though the correlation might seem rather low, indicating the already mentioned difficulty in distinguishing the two kinds of agents, their domination over the other ones appears to be clear.

Table 1. The variables most correlated with the labelling is_bot variable

Rank	Variable	isBot
1	isBot	1.00
2	avg_events_in_session_over_3_events_zscore	0.33
3	avg_events_in_session_over_3_events	0.33
4	uid_count_zscore	0.30
5	uid_count	0.30
6	max_distinct_page_views_per_hour	0.20
7	max_distinct_page_views_per_hour_zscore	0.20
8	max_distinct_page_views_per_minute	0.19
9	max_distinct_page_views_per_minute_zscore	0.19
10	csh_count	0.18
11	csh_count_zscore	0.18
12	csw_count	0.17
13	csw_count_zscore	0.17
14	page_view_cnt_zscore	0.17
15	traffic_share_zscore	0.17
16	traffic_share	0.17

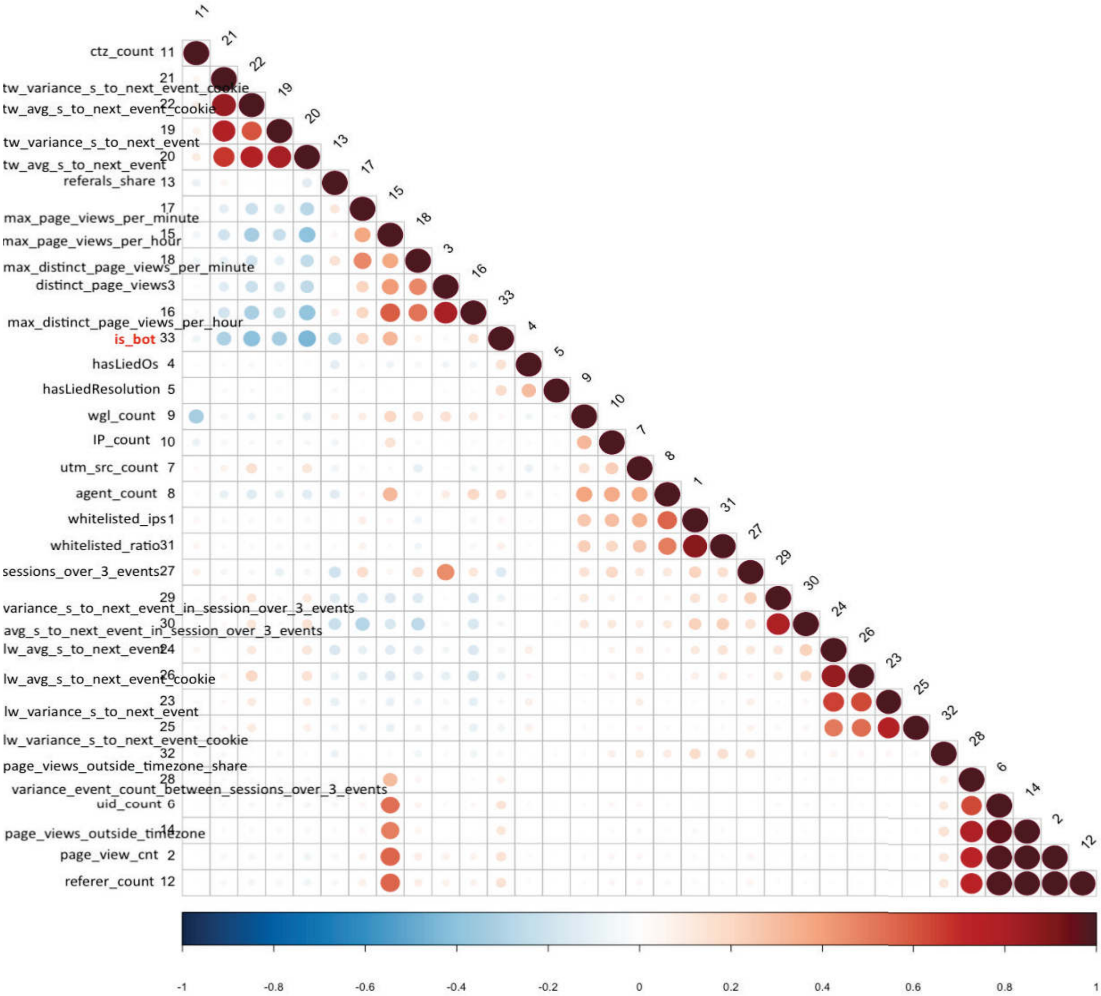
Yet, despite these observations, the choice of variables is by no means trivial nor unambiguous, and we shall return to this issue a bit later in this section.

Thus, we show, in Figure 6., another example of correlation analysis, performed for a smaller number of variables, namely 32, with the 33rd variable being the is_bot, which, as said, serves to establish the labels “bot” and “human”. Again, strong correlations in blocks are observed, meaning that – as in the previous case – the number of considered features might be significantly reduced.

Analysis of the correlations contributed to the selection of the following subset of features for some of the further analyses:

distinct_page_views, agent_count, max_distinct_page_views_per_hour, uid_count, max_page_views_per_hour, referals_share, variance_s_to_next_event_in_session_over_3_events, tw_avg_s_to_next_event, referer_count, avg_s_to_next_event_in_session_over_3_events, hasLiedResolution, max_page_views_per_minute, max_distinct_page_views_per_minute, tw_variance_s_to_next_event_cookie, page_view_cnt, tw_avg_s_to_next_event_cookie, tw_variance_s_to_next_event.

Figure 6. Correlogram for 33 features



It should be noted, though, that while the above selection of attributes was justified by the results from the correlation analysis reported, and, as could also be seen from the preceding example, the results from the correlation analysis were from this point of view quite clear, an important conclusion from numerous analyses performed was that the *overall stream of data considered displayed essential internal diversification*, and were a similar analysis carried out for a different segment of data, or the choice of attributes were based on a different methodology, another set of features would most probably result. One of the essential aspects of the diversification mentioned concerned the different advertising campaigns, with, apparently, the differences resulting both from the proper characteristics of a given campaign and the behavior of agents, either human or artificial. This observation seems to be crucial for the potential methodologies of coping with the here considered issue.

In this context, we show in the following two tables, Tables 2. and 3., two choices of variables, performed with the CFS method (1's in respective columns denoting variables effectively selected), for two different advertising campaigns and with application of a selection criterion, depending upon the completeness of respective data.

Table 2. Variables, selected by the CFS method for a data set, obtained from a campaign, depending upon the acceptance criterion of the variables

Variables selected	*Solely complete data*	*Data with gaps*
page_view_cnt	1	1
hasLiedResolution	1	1
max_page_views_per_minute		1
avg_events_in_session_over_3_events		1

Table 3. Variables, selected by the CFS method for a data set, obtained from a campaign, different from that for Table 2., depending upon the acceptance criterion of the variables

Variables selected	*Solely complete data*	*Up to 40% of missing*	*More than 40% of missing*
page_view_cnt	1	1	1
hasLiedResolution			1
max_page_views_per_minute		1	1
avg_events_in_session_over_3_events		1	1
referals_share	1	1	1
hasLiedOs	1		
tw_variance_s_to_next_event_cookie			1

3.3. Principal component analysis

Another approach to the analysis of the data available from the point of view of variable selection and interrelations between them, which was also tried out in this study, was the well-known method of principal component analysis (PCA).

PCA, similarly as the closely associated factor analysis, uses the matric of correlation coefficients to formally establish the most correlated variables and on their basis form a new set of variables, which are uncorrelated among them (a new, orthogonal system of coordinates). There are several types of output and conclusions that can be drawn from such an exercise:

(1) smaller number of aggregate variables, sufficiently well representing the original data;

(2) groups of original variables, forming the new ones – their composition and the (potential) interpretation of these new variables / their groups;

(3) explained variation of the original data by the consecutive new variables, and cumulatively, along with the shape of the respective curve (how well do we reconstruct the original data and at what point may we stop – see point 1 above);

(4) spatial image of the data in the new, orthogonal space – projections onto the two- or three-dimensional subspaces, which may very well illustrate the synthetic landscape of the data set.

The subsequent series of tables, treated together as Table 4. (i.e. Tables 4a, 4b, 4c and 4d) shows the results of the PCA for the data, concerning three different campaigns and jointly. The tables contain the so-called “loadings”, linked with respective correlations, and the way, in which the consecutive principal components (PCs) are formed by individual original variables. Note that we show here seven PCs for seven original variables, so that all the shown PCs ought to represent perfectly (100% of variance) the original variables (their variations).

Table 4. Loadings of the principal components (PC1 through PC7) for a series of datasets, corresponding to individual campaigns, denoted 1, 2 and 3, and jointly (colours show the direction and magnitude of loadings, associated with correlations)

Table 4a. Results for campaign no. 1

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7
page_view_cnt	-0.37	0.00	0.24	-0.70	-0.53	-0.13	0.06
hasLiedResolution	0.00	0.68	-0.22	-0.47	0.52	0.02	-0.01
max_page_views_per_minute	-0.54	0.01	-0.04	0.09	0.02	0.58	-0.60
avg_s_to_next_event_in _session_over_3_events	-0.54	0.02	-0.04	0.19	0.15	0.25	0.76
referals_share	0.00	-0.22	-0.94	-0.14	-0.23	0.00	0.03
hasLiedOs	0.02	0.70	-0.08	0.41	-0.58	-0.02	0.01
tw_std_min_to_next_event _cookie	-0.52	0.01	-0.09	0.23	0.18	-0.77	-0.22

Table 4b. Results for campaign no. 2

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7
page_view_cnt	0.46	-0.06	-0.32	0.04	0.04	-0.81	0.12
hasLiedResolution	0.01	0.66	-0.27	-0.04	-0.70	0.03	-0.01
max_page_views_per_minute	0.62	0.09	0.23	0.01	0.02	0.15	-0.72
avg_s_to_next_event_in_session_over_3_events	0.61	0.04	0.23	0.00	-0.04	0.35	0.67
referals_share	0.02	-0.17	0.06	-0.97	-0.13	-0.05	-0.01
hasLiedOs	-0.02	0.67	-0.13	-0.22	0.70	0.02	0.05
tw_std_min_to_next _event _cookie	0.16	-0.27	-0.83	-0.04	0.10	0.44	-0.06

Table 4c. Results for campaign no. 3

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7
page_view_cnt	-0.52	0.03	-0.24	0.25	-0.13	-0.65	0.40
hasLiedResolution	-0.02	0.69	-0.21	-0.19	0.66	-0.07	0.02
max_page_views_per_minute	-0.55	0.05	0.35	-0.17	0.00	-0.19	-0.71
avg_s_to_next_event _in_session_over_3_events	-0.49	0.07	0.45	-0.21	0.03	0.47	0.54
referals_share	0.19	0.08	0.61	0.69	0.30	-0.14	0.03
hasLiedOs	0.07	0.71	0.04	0.19	-0.66	0.12	-0.07
tw_std_min_to_next _event_cookie	-0.38	-0.07	-0.45	0.56	0.14	0.52	-0.21

Table 4d. Results for the three campaigns combined

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7
page_view_cnt	0.45	0.00	-0.23	-0.52	0.01	0.68	0.09
hasLiedResolution	0.01	-0.67	0.22	-0.08	-0.71	0.01	0.00
max_page_views_per_minute	0.56	0.10	0.35	0.05	0.01	-0.31	0.67
avg_s_to_next_event _in_session_over_3_events	0.60	0.05	0.20	-0.03	0.02	-0.25	-0.73
referals_share	-0.19	0.27	0.80	0.10	-0.02	0.48	-0.08
hasLiedOs	-0.01	-0.67	0.22	-0.04	0.71	0.04	0.01
tw_std_min_to_next _event_cookie	0.29	-0.14	-0.22	0.84	-0.03	0.38	0.00

Just by looking at the results from Table 4. one can gain quite important insights into the significant features of the data here considered. First, as already noted, even if there is quite a high degree of qualitative agreement as to the components determined among the campaigns, there are also distinct differences, which are also strongly expressed by the differences with respect to the results for the combined data set. Second, the interpretation of the particular components is in a vast majority of cases quite obvious, conform to what the common sense – and the expert opinion – would suggest. Such conclusions, though, are largely dependent upon the scale of variation explained by the consecutive CPs and cumulatively. This, in turn, is the subject of Table 5., structured analogously to Table 4.

Table 5. Shares of variance explained by particular components in the PCA performed

Table 5a. Results for campaign no. 1

Items	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.71	1.16	1.01	0.84	0.79	0.49	0.38
Share of explained variance	0.42	0.19	0.14	0.10	0.09	0.03	0.02
Cumulative share of variance	0.42	0.61	0.76	0.85	0.95	0.98	1.00

Table 5b. Results for campaign no. 2

Items	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.44	1.09	1.02	1.00	0.91	0.81	0.45
Share of explained variance	0.30	0.17	0.15	0.14	0.12	0.09	0.03
Cumulative share of variance	0.30	0.47	0.62	0.76	0.88	0.97	1.00

Table 5c. Results for campaign no. 3

Items	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.43	1.13	1.06	0.95	0.84	0.77	0.61
Share of explained variance	0.29	0.18	0.16	0.13	0.10	0.08	0.05
Cumulative share of variance	0.29	0.47	0.63	0.76	0.86	0.95	1.00

Table 5d. Results for the combined data set

Items	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.48	1.15	1.04	0.97	0.84	0.73	0.48
Share of explained variance	0.31	0.19	0.15	0.13	0.10	0.08	0.03
Cumulative share of variance	0.31	0.50	0.66	0.79	0.89	0.97	1.00

The results of Table 5. bring some important observations, which are further corroborated by the visual illustrations of these results, provided in Figure 7a, b. Thus, Campaign no. 1 appears to be distinctly different from the other ones in that in its case four components seem to suffice for the adequate representation of the original data, while in the remaining cases five components seem to be necessary. This is primarily due to the very high share of variance, explained by the first component (42%).

Figure 7a. Illustration to Table 5. Series correspond to campaigns and their joint consideration

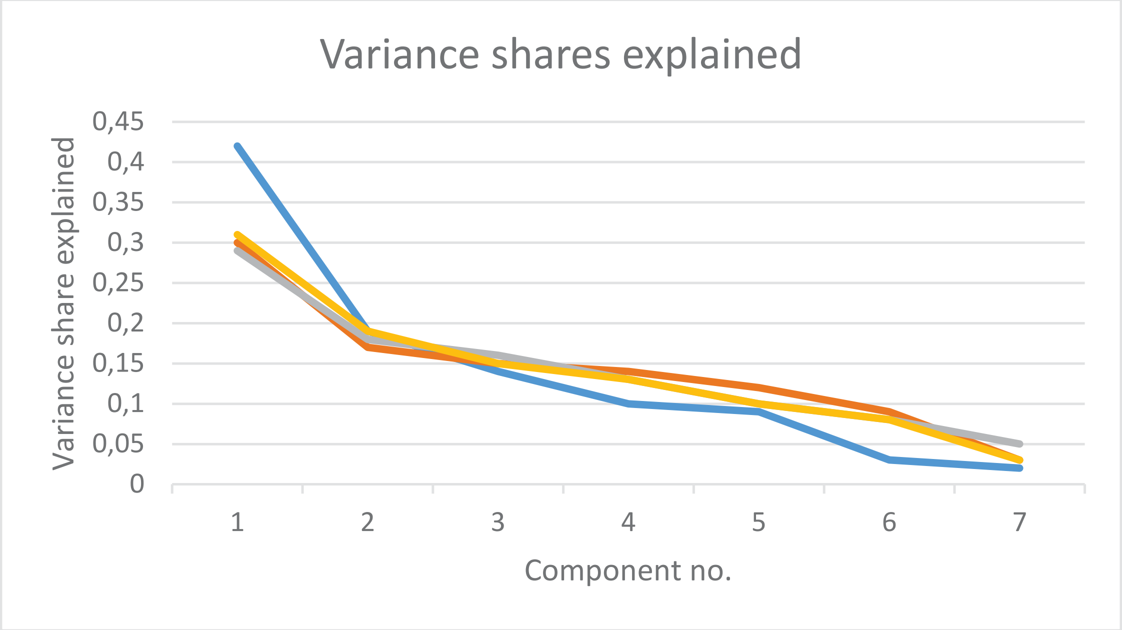
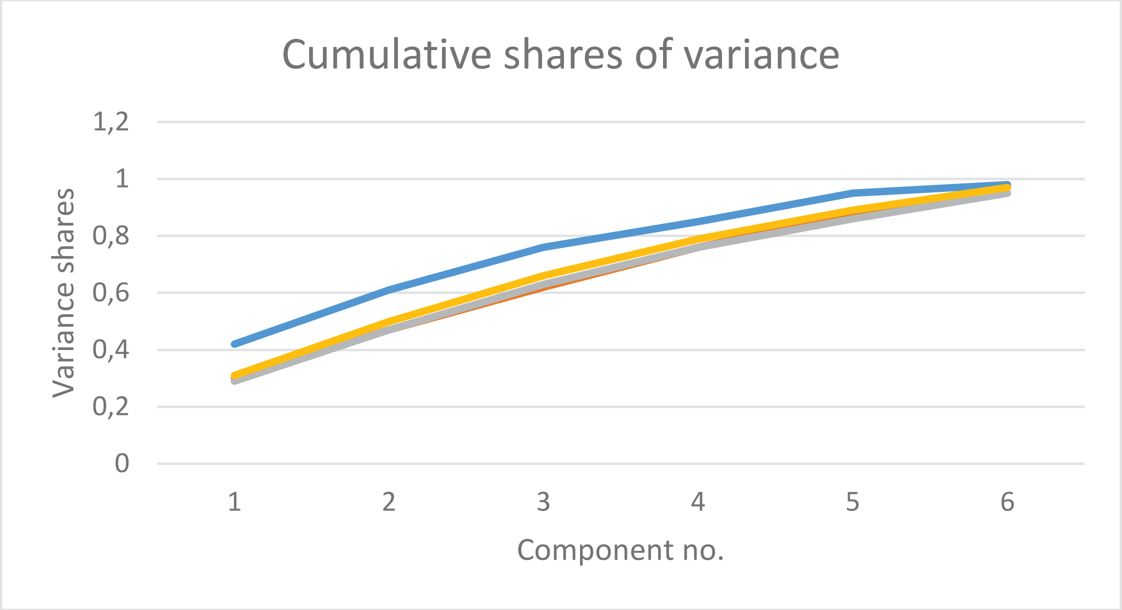


Figure 7b. Illustration to Table 5. Series correspond to campaigns and their joint consideration



Yet, despite this difference, the relatively similar composition of the two first principal components across all campaigns analysed appears to bring some hope in terms of the possibility of distinguishing the two kinds of agents considered. This hope, though, is not in any way confirmed by the projections, which are shown in the following series of figures.

Figure 8a. Projections on the main component planes for Campaign no. 1

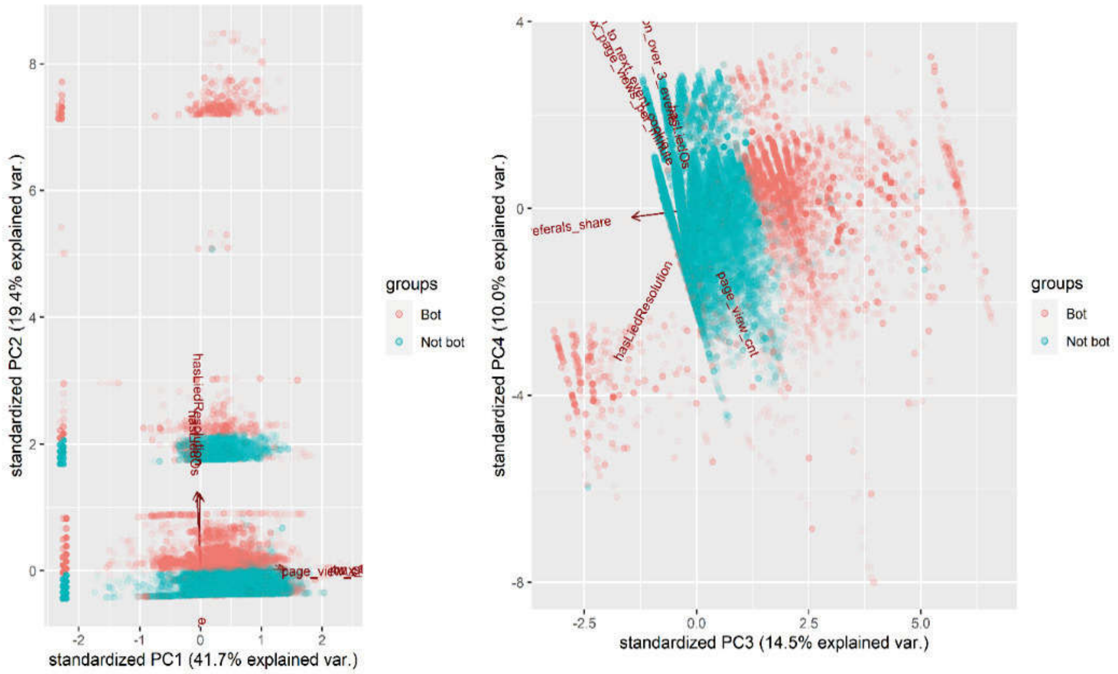


Figure 8b. Projections on the main component planes for campaign no. 2

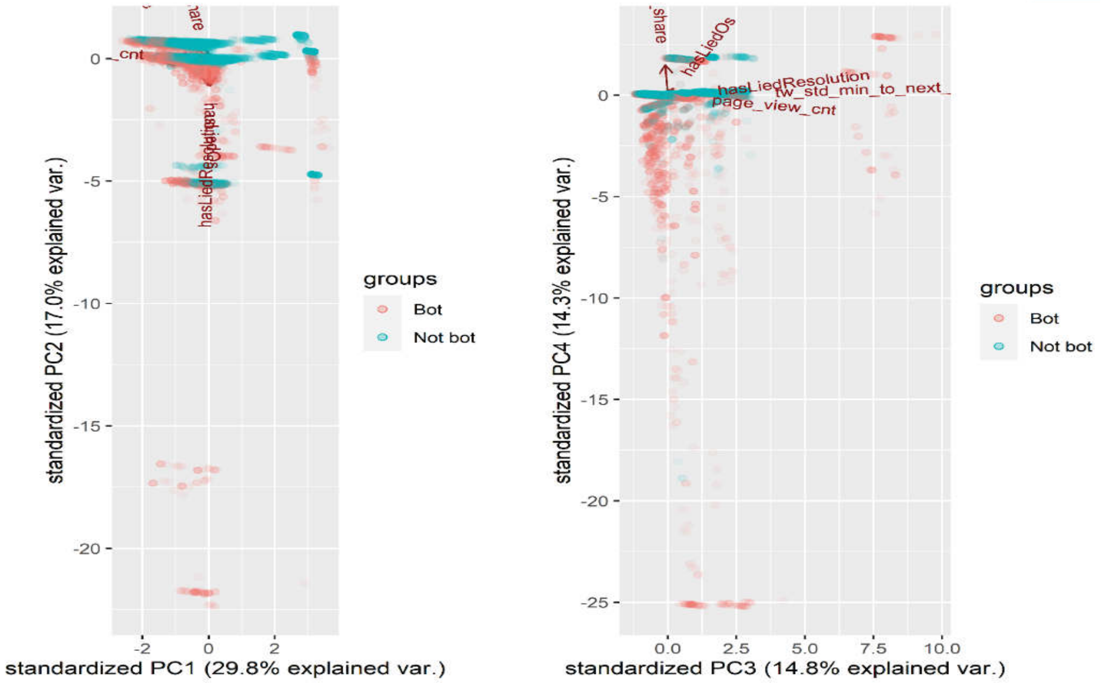


Figure 8c. Projections on the main component planes for campaign no. 3

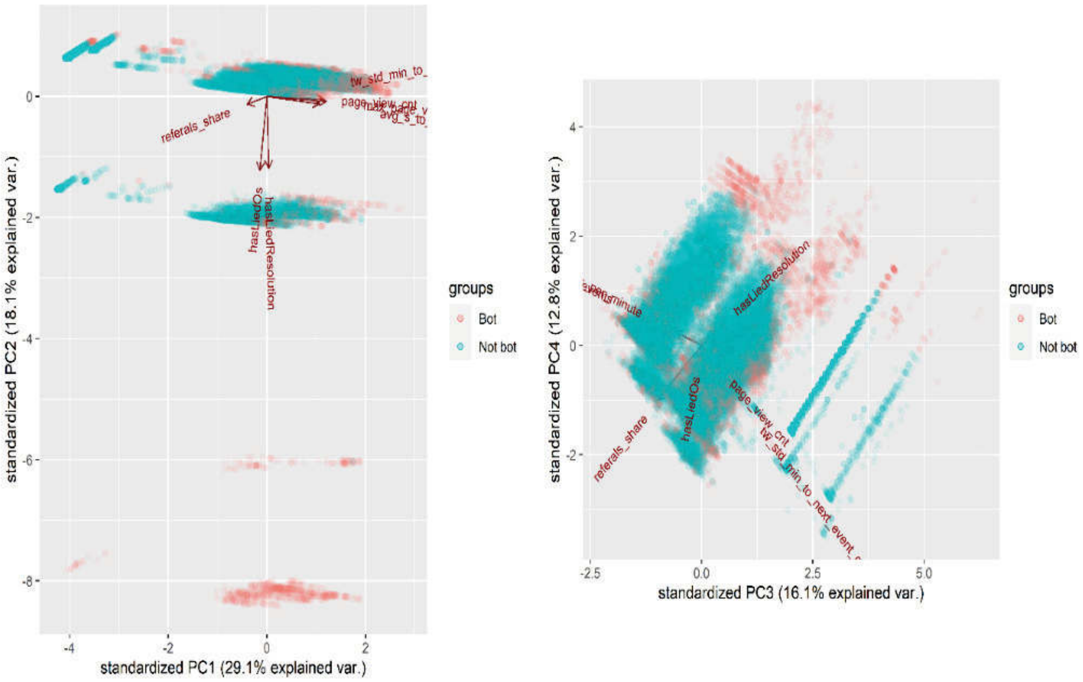
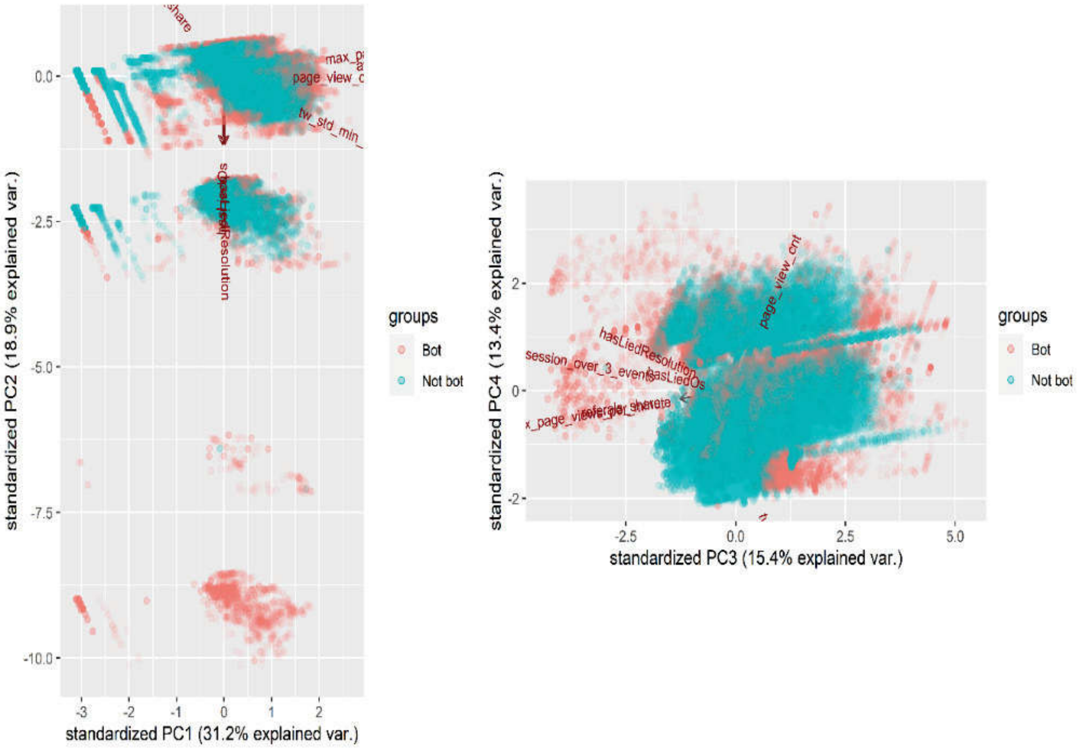


Figure 8d. Projections on the main component planes for the data for all three campaigns

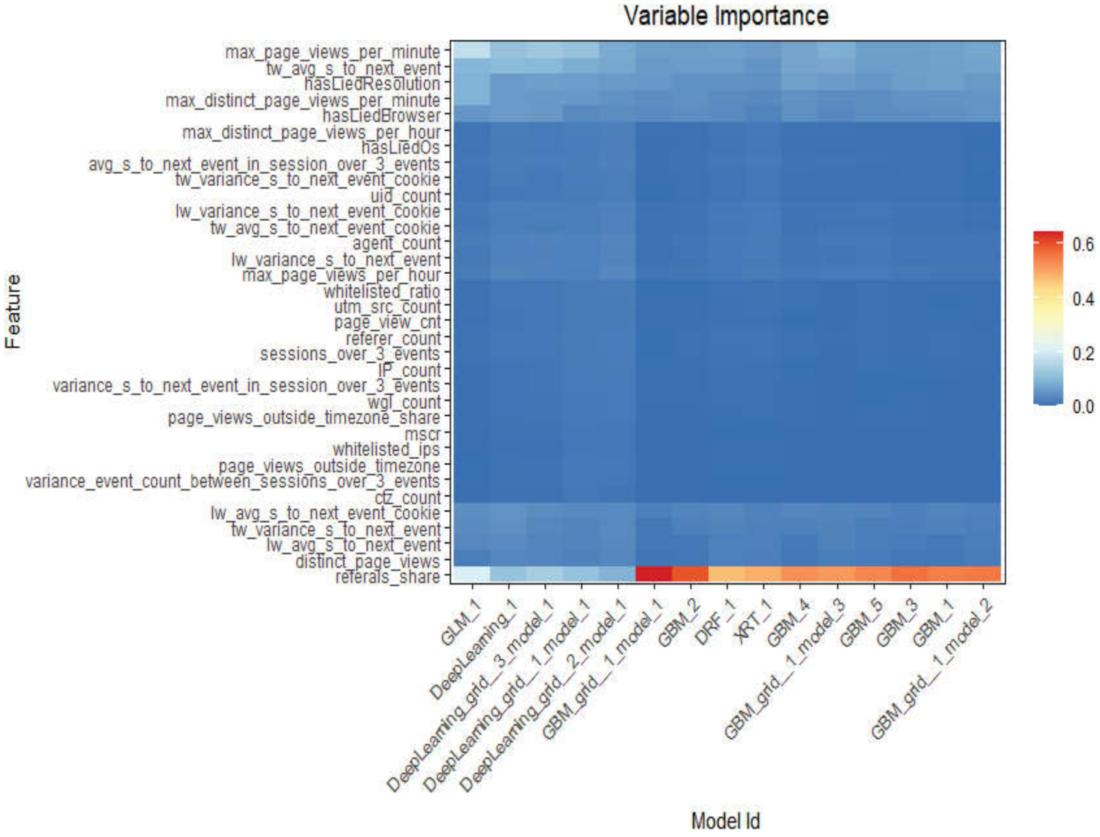


Now, these figures explain a lot, concerning almost all of the previous observations. Thus, we can now summarise the findings from this part of the study, devoted, in principle, to the question of variables and their selection, but, actually, having also an important bearing on the understanding of the entire image of the data analysed. It turns, namely, obvious from these figures that we deal with a number of subsets of observations (several of them), some of them containing exclusively either bots or humans, but, to a large extent, also a mixture of the two. Further, it appears that the image (especially concerning the mixed groups) is simpler for the individual campaigns than for the data sets from different campaigns, strengthening the already formulated conclusion of the differences between the campaigns.

3.4. Variable importance according to h2o package

In the framework of the analysis of variables and their potential role in the data set considered yet another attempt was done, namely of applying h2o package, which produces the heatmap of the variables, generated by the appearance and weight of particular variables in the automatically applied AutoML (machine learning) techniques. The result for a definite set of variables is shown in Figure 9.[8]

Figure 9. Results of application of the h2o package



[8] See Model Prediction Correlation Heatmap — h2o.model_correlation_heatmap • h2o

What is highly characteristic of this result is that the approach indicates only one variable as really "hot", i.e. referals_share (although not across the entire spectrum of techniques). We have already seen that this variable was indicated as one of the most important in the preceding exercises (see, for instance, Table 4.), but it appeared as important along with some of the other ones. Here, it stands in a way alone. This is most probably due to the fact that the other variables, otherwise appearing as important, can be replaced by yet other ones, which are closely associated with them (different kinds of aggregates of a similar quantity).

3.5. Comparison with the existing blacklists

Although this specific analysis is only loosely associated with variable analysis, it definitely belongs among the introductory stages of work, meant, first of all, to shed some general light on the characteristics of the observed events and behaviours. In this sense, it can corroborate the results from the analysis of variables or can indicate some special directions of search.

Thus, a comparison was performed of the labelling by the rule-based expert tool, used in this study with the spammer blacklists, available on the web (RBL, *Real-time Blackhole List*). In this case, two services were used in the comparison, namely http://multirbl.valli.org and https://mxtoolbox.com/. Table 6. shows an excerpt from this comparison. Rows correspond to exemplary IPs, which are not provided here, while columns – to the indication of the own rule-based tool, and the output from the two services, in these two columns the entries having the form of x/y, where x is the number of lists, on which the given IP was found, and y – the total number of lists checked. The numbers of lists may vary (in the case of http://multirbl.valli.org) in view of dependence on response from the rDNS.

One can see in Table 6. that there are cases, in which the indication from our tool finds a consensual confirmation in terms of the blacklists (see first four rows of the table), but also quite a number of cases, where there are inconsistencies, both between our tool and the blacklists, and between the blacklist sources (there was no attempt made to select the cases here presented especially in order to show these inconsistencies). Thus, for instance, the indication of our tool of a "whitebot: a crawler" is confirmed by the result for https://mxtoolbox.com/, but not so for the other service. Finally, at the end of the here quoted set of cases, several non-indications by our tool are given, which are blacklisted according to both of the services referred to.

The ultimate conclusion from this exercise, here only shortly illustrated in order to show the most important features of the outcome, is that even if the existing blacklisting information might be of some use in the case of the problem here considered, it ought to be treated either as only an auxiliary information, or as one of the elements of the pertinent set of characteristics.

This, definitely, confirms the already formulated proposition that the problem at hand is quite complex and cannot be solved in a straightforward manner by the use of some well-known method, to be just 'taken off the shelf'.

Table 6. An excerpt from the comparison of indications of the rule-based tool and the blacklisting services

Own tool used in ABTShield	**Blacklists conform to https://mxtoolbox.com/**	**Blacklisty conform to http://multirbl.valli.org**
Bad Bot: Impersonator	3/87	17/243
Bad Bot: Impersonator	3/87	22/243
Bad Bot: Impersonator	4/87	18/243
Bad Bot: Impersonator	2/87	13/243
Bad Bot: Impersonator	0/87	8/243
Bad Bot: Impersonator	0/87	4/189
Bad Bot: Impersonator	0/87	4/243
Bad Bot: Impersonator	4/87	13/243
Bad Bot: Third party	0/87	3/189
Bad Bot: Third party	0/87	1/243
Bad Bot: Third party	2/87	15/243
Whitebot: crawler	0/87	5/243
Whitebot: crawler	0/87	5/243
Whitebot: crawler	0/87	5/243
Whitebot: crawler	0/87	5/243
NULL	1/87	16/243
NULL	2/87	13/243
NULL	1/87	11/243
NULL	1/87	7/243
NULL	3/87	13/243
NULL	3/87	15/243

3.6. Conclusions from the study of variables

The general conclusions, of highest importance for the further work, that we can formulate at this stage, i.e. in the consequence of having performed the analyses here reported, are:

- The distinction between bots and humans is by no means a simple construct, like, e.g., a single discriminating rule or a couple of condition.
- This is largely due to the fact that there are, evidently, more than two (i.e. “bot” and “human”) behaviour patterns (groups, subsets, clusters), as expressed by the variables accounted for. These behaviour patterns may correspond uniquely to bots or humans, but, apparently, there are also such ones that characterise both kinds of agents and distinction of these kinds inside them is certainly difficult.
- It is possible to characterise relatively well the overall variation of the data with a low number of variables (just a couple of them), but this does not guarantee a high precision in the attempts to distinguish people from bots, and so the issue of best choice of variables remains important.

- It appears that the methods, meant to distinguish people and bots ought to be tuned to particular campaigns (or other substreams of data) in view of the fact that there are quite significant differences between them.

4. The approaches tested

The further work on the problem, that is – the search for an effective method of distinguishing the bots from humans, which would either surpass the expert-designed rule-based querying tool, or at least be on par with it, while providing responses in a more computationally efficient manner, involved a broad array of methodologies. This concerned, first, the choice of attributes, commented upon before, and then identification (classification, categorisation) of observations. With respect to the latter, appearing to be of foremost importance, the techniques that were tried out include clustering (a couple of diverse clustering algorithms), reverse clustering (see Owsiński et al., 2021), identification of association rules, as well as diverse classifier building techniques. A special effort oriented at the comparison of results, related to the latter, was undertaken, in which such techniques were applied as, in particular, those from the WEKA toolset, namely bagging, random forest, LMT (using logistic regression), or JRip (using rule system). Regarding these techniques, also synthetic treatment (e.g., voting) was tried out.

Some hybrid or combined approaches were tested as well in the framework of this research, consisting in the conjoint application of clustering and classifier training, i.e. partitioning of the set of observations into subsets and then training of the classifiers proper for the subsets. This general concept was implemented in a certain variety of manners, with, for instance, not only tuning of parameters and selection of variables corresponding to different subsets, but with selection of different techniques for the particular subsets, or even abandoning of classifier identification if the subsets obtained were (sufficiently) "clean" (i.e. all elements, or a definite proportion of them belonged to either human or bot categories). The subsequent two sections of the paper report on the selected attempts, out of those mentioned above, starting with straightforward clustering approaches.

5. Clustering of observations

Clustering, as this is well known, consists in grouping of similar observations, while separating the dissimilar ones. There are lots of kinds of clustering algorithms and of their variations in each of these kinds (hierarchical aggregation algorithms, k-means-like algorithms, density-based algorithms, etc.). Each of these kinds of algorithms operates with definite parameters, which have to be specified prior to the execution of the algorithm on a given set of data. In addition, it is also known that different kinds of clustering algorithms function with different effectiveness for various data sets, meaning such characteristics as the (hypothetical) numbers of clusters, shapes of clusters, dimensionality of the respective feature space, measurement scales of variables, representing observations, etc.

5.1. An initial exploratory study

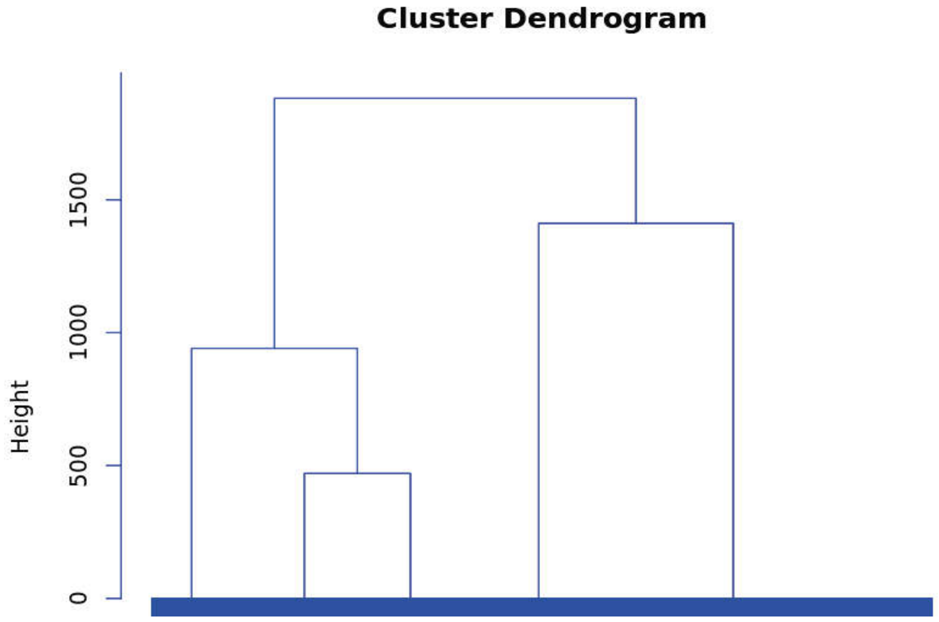
A number of relatively small scale attempts were first made in order to:

(a) check the capacity of clustering in the framework of the problem considered; and

(b) possibly determine the direction of search, concerning the most appropriate clustering procedure for the problem at hand.

We shall report in this subsection on one of such attempts, based on a hierarchical aggregation algorithm. The algorithm of choice was Genie (see Gagolewski et al., 2016), a fast and robust algorithm from this family. Figure 10. shows the sketch of the dendrogram[9], produced with this algorithm on the basis of relatively small sample of 40 thousand observations, characterized by the variables, selected with the CFS method, with, what is important, balanced representation of "humans" and "bots".

Figure 10. Sketch of the dendrogram produced by the Genie algorithm for a sample of observations



This illustration is limited to the four final mergers, performed by the algorithm, showing five clusters as the starting point, rather than all of the 40 thousand observations. This is on purpose, since we were interested, in particular, in the "most appropriate" number of clusters, as well as, of course, their composition. In the case of hierarchical aggregation algorithms an external criterion is used to determine the level of the dendrogram, at which the "proper" solution is established, this criterion

[9] A dendrogram is a graphical illustration of the mergers of clusters along the functioning of the hierarchical merger algorithms, starting with all observations being separate clusters and ending with all of them forming one all-embracing cluster.

being usually of a statistical character. For our study, we used another criterion, namely we took the labels of "bot" and "human" as determining a division of the set of observations into two subsets and looked for the partition of the same set, obtained with the hierarchical aggregation algorithm used, Genie, that would be most similar to the division into "bots" and "humans". We used the classical rand index for this comparison.

The Rand index counts the pairs of observations, say, *n* observations, that is – ½ *n*(*n*-1) of pairs, in the following manner:

# those that are in the same cluster in both partitions considered (*a* pairs), or

# in different clusters in the both of the two partitions considered (*b* pairs), or

# the pair is in the same cluster in one partition and in separate clusters in the second (*c* pairs), or, finally,

# the pair is in separate clusters in one partition and in the same cluster in the second partition (*d* pairs).

Of course, ½ *n*(*n*-1) = *a* + *b* + *c* + *d*. With these counts, the "raw" Rand index is equal (*a*+*b*)/(½ *n*(*n*-1)), when it is maximized (search for maximum similarity of the partitions, taking values between 0 and 1. It can be used also in its minimized form, and is often complemented with a correction for the statistical bias, by deducting the number of pairs that can be in agreement randomly.

The run of the algorithm that is illustrated in Figure 10., produced the Rand index values as shown in Figure 11. The Rand index used in this exercise was corrected for the bias mentioned.

This curve indicates that the best similarity of the two partitions considered (i.e. the one into two groups according to labels and the one obtained from the hierarchical aggregation Genie algorithm) occurs for the partition into four clusters according to the results from Genie (the value of the Rand index ought not necessarily be the "best" for the same number of clusters in the two compared partitions). It should be noted, though, that the values of the Rand index are very low indeed, actually one can think whether they are at all significant and whether they do not indicate lack of possibility of obtaining (with the use of Genie) of the results that are really somehow similar to those implied by the labelling.

This supposition is, indeed, confirmed by the content of the obtained four clusters, shown in Figure 12. The image illustrates the shares of "bots" and "humans" (or "not-bots") in particular clusters, forming the partition, obtained with the Genie algorithm. The result is definitely disappointing.

Figure 11. Values of the corrected Rand index for the run of the Genie algorithm commented upon in the text

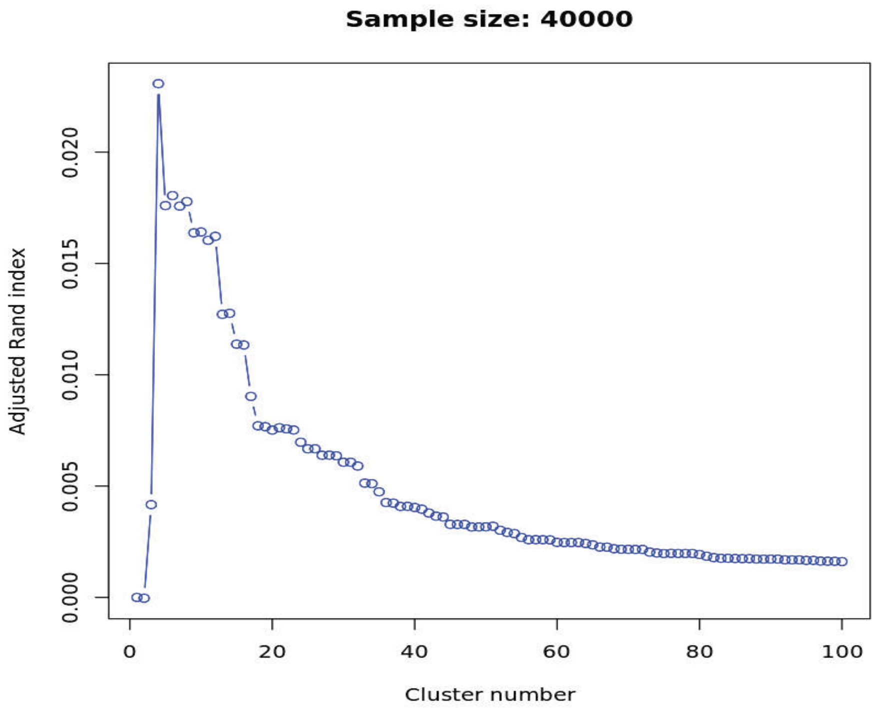
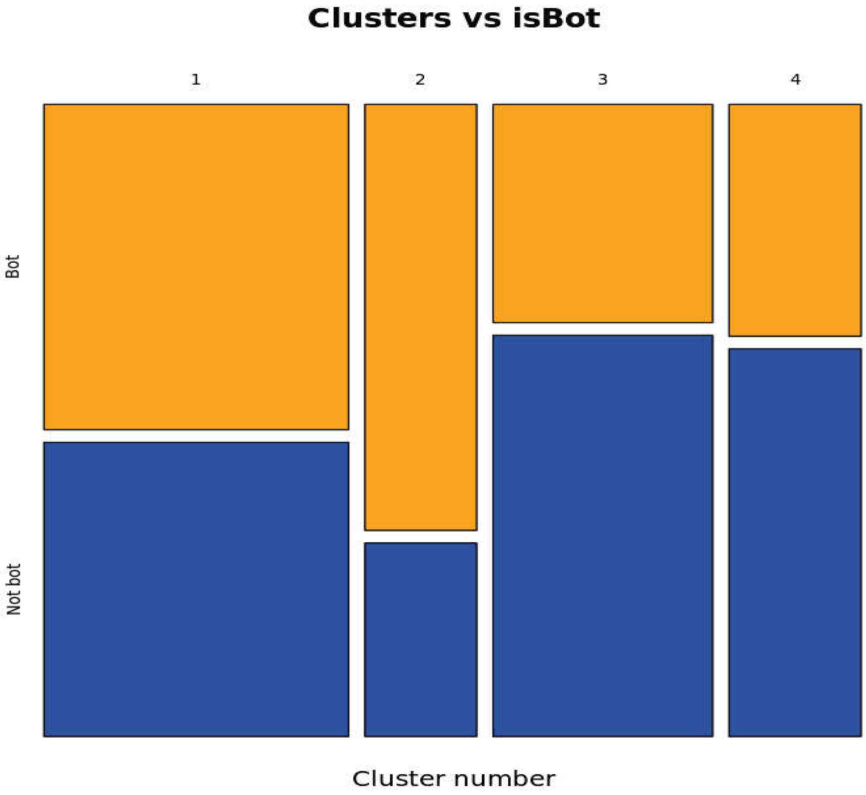


Figure 12. Content of the four clusters, determined by the Genie algorithm, against the labelling with the is_bot variable ("bots" in orange)



It appears, from Figure 12., as if the two kinds of observations were assigned to the clusters in a totally random manner, meaning that there are no regularities in the spatial structure of observations, indicating the dominance of any of these two types. At this stage it is really hard to say where the reason lies – in the algorithm itself, the variables used, the balancing of the sample…?

Since we were of the opinion that some way of using clustering should provide a valuable insight into the data we analysed, a number of different exercises were yet undertaken.

5.2. Temporal stability of clusters

This exercise differed from the preceding one both in terms of its purpose, explained in the title of this subsection, and in the clustering algorithm employed. In this case we used one of the most popular density-based algorithms, DBSCAN. This algorithm finds locally dense groups of observations and processes them to find the ultimate partition (see Ester et al., 1996; or, perhaps, Ling, 1972, for a supposedly earlier invention).

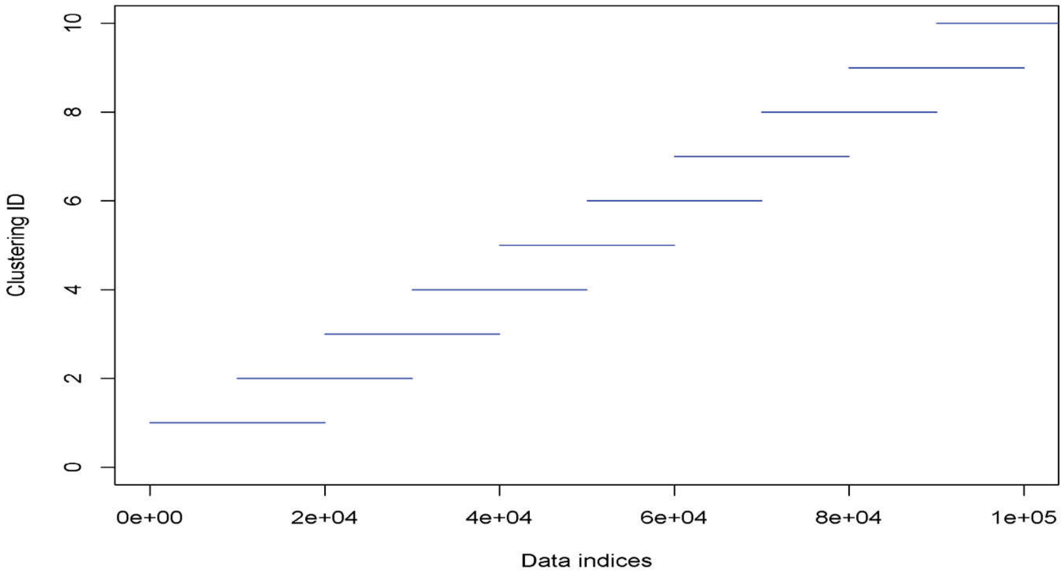
Like in many other density-based clustering algorithms, also DBSCAN, in its various implementations, uses two basic parameters, which are, in fact, decisive for the nature of the results obtained, namely a certain distance parameter ("neighbourhood radius"), defining, actually, what is considered to be "near", and a cardinality parameter ("magnitude of the neighbourhood"), which allows for the control of density. Most implementations provide default values or allow the user to set own values of these key parameters.

The experiment concerned the observations, characterized, very much like in the preceding one, by the variables, indicated by the correlation analysis, namely:

1. avg_events_in_session_over_3_events_zscore,
2. uid_count_zscore,
3. max_distinct_page_views_per_hour_zscore,
4. csh_count_zscore,
5. lw_avg_s_to_next_event_cookie_zscore,
6. utm_src_count_zscore,
7. tls_count_zscore,
8. tcp_count_zscore.

The data set consisted of altogether of 244 thousand observations for a single advertising campaign, due to the care for the supposed uniformity of data character. The set, which encompassed 23 consecutive days of the campaign, was split into temporal samples, moving by a single day to include new observations, with two important principles preserved: (i) the observations were (again!) balanced as to the content of the supposed "bots" and "humans"; (ii) each next subsample differed from the preceding one by exactly half of observations. This rule is illustrated in Figure 13., showing the sliding window of the samples and also their temporal span, equivalent to their time-wise content.

Figure 13. The temporal structure of the samples analysed in the study of the time-wise stability of clusters, generated with DBSCAN



The similarity of the consecutive partitions, established with DBSCAN for the consecutive subsamples, was (again!) measured with the corrected Rand index. The consecutive subsamples contained 10 000 balanced observations each. The clustering was performed for 20 consecutive sliding window subsamples. The comparisons were made with respect to the results, obtained from DBSCAN for the fourfold sample of 40 000 balanced ultimate observations, intended to represent a "prediction".

The results, regarding the behavior in time of the partitions obtained, depending upon the "distance in time", are shown in Figure 14. This diagram shows the values of the corrected Rand index for the partitions, featuring increasing "distance in time", represented by the horizontal axis of "Difference between clustering IDs". Naturally, the numbers of partitions compared for the increasing "distance in time" decreases, down to just a single one.

It can be concluded, on the basis of this illustration, that the stability of partitions over time is quite high, especially if we consider that the values of the Rand index are (also) relatively high (which sheds additional light on the results from the preceding subsection). The decline with time is not only slight, but, in fact, ambiguous.

Figure 14. Rand index values for the partitions with increasing temporal distance between them (DBSCAN, distance in days)

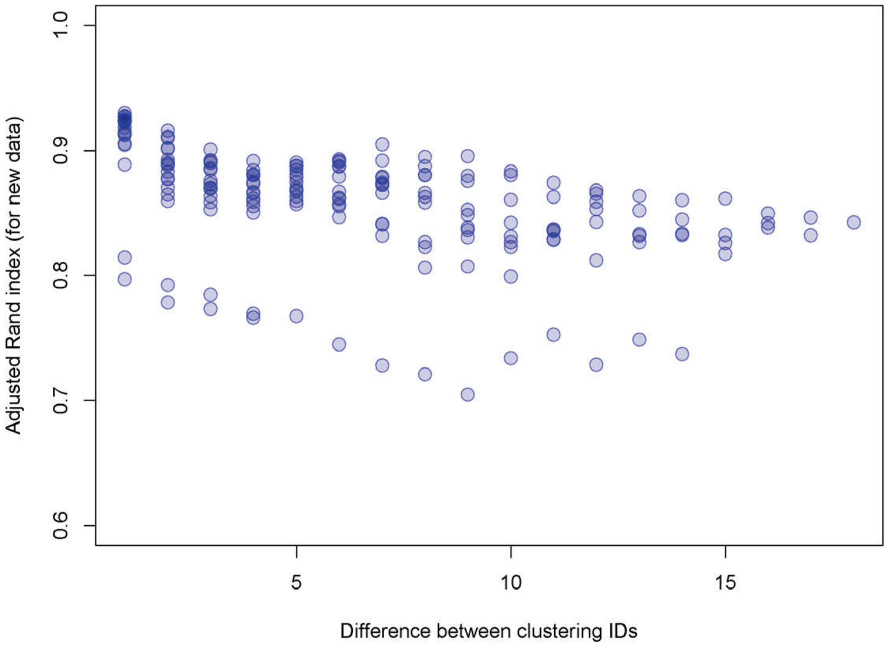
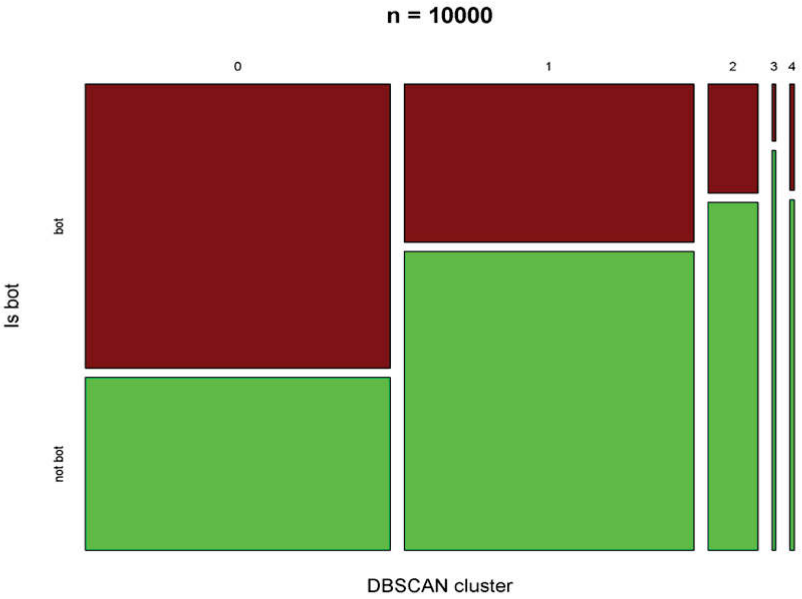


Figure 15. Average composition of the partitions generated by DBSCAN



Yet, a look at Figure 15. suggests, again, that, despite the relative stability, implying the possibility of obtaining also quite stable clusters along time, the results are not very promising, to say the least. Out of five clusters that were usually obtained, none showed a clear dominance of "bots", and only two marginal ones displayed a relatively important majority of "not-bots". This is, indeed, not a very good prognostic for the assumed subsequent experiments with the clustering algorithms.

5.3. The 'reverse clustering' approach

The next exercise was performed with the use of the 'reverse clustering' approach, as described in detail in Owsiński et al. (2017, 2021). The approach consists in solving of the following general problem:

Having a data set (set of observations) and some partition of this set into subsets, find (the parameters of) a clustering procedure that produces, for the given data set, a partition that is possibly similar to the one that is given.

This amounts to an optimization problem, consisting in the minimization of difference (or maximization of similarity) between two partitions, one, which is given, and another one, that is generated by the clustering procedure. Optimisation is performed with respect to the vector of parameters, determining the entire clustering procedure, that is:

- the choice of the clustering algorithm itself (e.g. one of the hierarchical aggregation algorithms, one of the density-based algorithms, a k-means-type algorithm,…)
- the setting of the algorithm-proper parameters (e.g. the values of the Lance-Williams coefficients for the hierarchical aggregation algorithms, the number of clusters for the k-means, neighbourhood radius and neighbourhood magnitude for density-based algorithms, etc.)
- the assumed definition of distance between the observations (e.g. the general Minkowski distance with variable exponent)
- the selection or the weights of the variables, describing the observations.

The criterion, used to assess the quality of solutions, may, for instance, be the already used here Rand index, whether in its raw, or corrected form.

The problem, as stated in this manner, is quite cumbersome. Not only the vector of the optimized parameters changes from one clustering algorithm to another, but the character of these optimization variables is diverse (discrete, continuous, binary…), the shape of the criterion function "landscape" is very complex, etc. That is why it is most frequent to use evolutionary algorithms to perform this optimization, although comparisons with other kinds of optimization algorithms were also carried out.

In the here presented exercise the following specifications were applied:

The selection of variables:

1. page_view_cnt
2. hasLiedResolution
3. max_page_views_per_minute
4. avg_events_in_session_over_3_events
5. referals_share

6. hasLiedOs
7. tw_variance_s_to_next_event_cookie

A number of optimisation algorithms was tried out, but we shall present here the results for the self-adaptive differential evolution algorithm. The clustering algorithms accounted for were: k-medoids ("pam"), and two hierarchical aggregation algorithms – Genie, already mentioned here, and one of the classical algorithms ("agnes"). The sample considered contained 10 thousand observations and was balanced regarding "bots" and "humans". All non-binary variables were normalized via logarithmic transformation.

Table 7. provides the results, concerning primarily the weights of variables, obtained in this exercise, along with all other pertinent information.

Table 7. Results from an exercise with the 'reverse clustering' approach. Parameter k is the number of clusters, p is the Minkowski exponent, while α | λ define the hierarchical aggregation algorithm

Data set	Algo-rithm	Rand index	Variable no.*							Parameters		
			1	2	3	4	5	6	7	k	p	α | λ
1	genie	0.15	0.351	0.206	0.195	0.441	0.179	0.279	0.170	10	0.455	0.227
	pam	0.108	0.120	0.114	0.102	0.121	0.112	0.118	0.122	9	2.646	
	agnes	0.1	0.276	0.187	0.203	0.295	0.272	0.171	0.119	7	0.784	3.952
2	genie	0.125	0.119	0.110	0.111	0.110	0.120	0.107	0.107	4	0.113	4.373
	pam	0.223	0.120	0.117	0.133	0.117	0.110	0.131	0.134	3	0.69	
	agnes	0.212	0.150	0.207	0.184	0.167	0.250	0.168	0.241	3	0.9	0.516
3	genie	0.108	0.153	0.133	0.186	0.147	0.146	0.143	0.115	5	0.107	4.664
	pam	0.091	0.121	0.106	0.124	0.123	0.125	0.117	0.117	7	1.44	
	agnes	0.06	0.173	0.164	0.396	0.186	0.337	0.195	0.124	5	0.999	2.149

Average variable weight (Rand index weighted)	0.174	0.152	0.168	0.189	0.175	0.159	0.149

These results, even if significantly better than those from subsection 5.1 (see the values of the Rand index) are still quite far from being satisfactory. They confirm, though, high diversification of the data sets among the campaigns, as well as existence of more than just two categories in the data (clusters). What is also of some interest is that in the selected set of variables all seem to have more or less equal importance.

5.4. Extended analysis with k-medoids algorithm

We shall present now a more elaborate analysis, performed with just one clustering algorithm, namely k-medoids ("pam"), using various data sets and additional operations. The variables, which were used in this series of experiments were as follows:

1. is_bot – the labelling variable, here in an extended form, taking values from 0 ("not-bot") upwards, indicating increasingly strong suggestion that an observation is a "bot"
2. max_page_views_per_minute
3. tw avg s to next event cookie
4. distinct page views
5. max distinct page views per hour
6. lw avg s to next event cookie
7. max distinct page views per min
8. cnt of cols where z score exceeded 5
9. wgl count
10. lw min s to next event cookie
11. variance events in user Agent session
12. lw variance s to next event
13. tls count
14. tw variance s to next event cookie
15. sessions over 3 events per allcnet and user Agent

The algorithm was applied to a series of data sets from various campaigns, most often of some 20 000 observations each. In distinction from what we have seen in the preceding examples, in this case a definite success was achieved, as we shall see further on, which may be attributed to: (a) the scaling of the labelling variable (taking discrete values spaced by 0.25 from 0 upwards), (b) the starting assumption of a bigger number of clusters (determined, effectively, with the silhouette approach), and (c) lack of balancing of the samples (proportions of the elements with various is_bot values were not subject to any selection).

Thus, Table 8. shows an example for a run, in which nine clusters, corresponding to columns, were obtained, with, in consecutive rows, characterisations of the clusters in terms of the numbers of observations, contained in each cluster, featuring definite values of the is_bot variable, starting from 0 ("not-bot" or "human"), up to as far as 9.5 ("certainly a bot").

It is, definitely, striking, that in this case we obtained, first, certain totally "pure" clusters, namely: clusters, containing only "bots" (nos. 1, 5 and 9), and clusters, containing only "not-bots" (nos. 2, 3, 4 and 7). The remaining two clusters, that is – nos. 6 and 8 – contain roughly 0.1% of "bots", virtually all their elements being "not-bots".

Table 8. An instance of results obtained with the "pam" k-medoids algorithm. Entries are the numbers of elements of individual clusters, characterized by respective values of is_bot variable

is_bot degree	**Clusters** and numbers of their elements								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	0	2464	4737	3109	0	1428	1177	898	0
1.25	360	0	0	0	0	11	0	9	1235
1.5	46	0	0	0	0	0	0	1	62
1.75	362	0	0	0	0	1	0	0	116
2	111	0	0	0	1	0	0	0	3
2.25	371	0	0	0	2	1	0	0	4
2.5	38	0	0	0	22	0	0	0	0
2.75	24	0	0	0	139	0	0	0	0
3	0	0	0	0	37	0	0	0	0
3.25	0	0	0	0	131	0	0	0	0
3.5	0	0	0	0	51	0	0	0	0
3.75	0	0	0	0	65	0	0	0	0
4	0	0	0	0	16	0	0	0	0
4.25	0	0	0	0	25	0	0	0	0
4.5	0	0	0	0	29	0	0	0	0
4.75	0	0	0	0	23	0	0	0	0
5	0	0	0	0	14	0	0	0	0
5.25	0	0	0	0	8	0	0	0	0
5.5	0	0	0	0	6	0	0	0	0
5.75	0	0	0	0	3	0	0	0	0
6	0	0	0	0	5	0	0	0	0
6.25	0	0	0	0	2	0	0	0	0
6.5	0	0	0	0	5	0	0	0	0
6.75	0	0	0	0	3	0	0	0	0
7	0	0	0	0	4	0	0	0	0
8	0	0	0	0	1	0	0	0	0
9.5	0	0	0	0	1	0	0	0	0

The fact that virtually "pure" clusters were obtained is not the only striking feature of this result (and a lot of similar results). Namely, if we look at the three "bot" clusters, we clearly see that they are distinctly different: There are two bigger clusters, 1 and 9, which contain observations, characterized by not too high indications of "botness", this especially applying to cluster 9, but cluster 5 obviously groups observations, which are definitely "bots". In order to have a comparative material, we quote yet another analogous result in Table 9.

Table 9. Another instance of results obtained with the “pam” k-medoids algorithm. Entries are the numbers of elements of individual clusters, characterized by respective values of is_bot variable

is_bot degree	**Clusters** and numbers of their elements									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	3493	3664	4	2593	1146	848	0	1608	2401	0
1.25	0	0	567	0	3	10	0	1	0	895
1.5	0	0	47	0	0	0	0	0	0	67
1.75	0	0	375	0	0	0	6	0	0	238
2	0	0	27	0	0	0	41	0	0	24
2.25	0	0	34	0	0	0	218	0	0	27
2.5	0	0	0	0	0	0	76	0	0	0
2.75	0	0	0	0	0	0	185	0	0	0
3	0	0	0	0	0	0	54	0	0	0
3.25	0	0	0	0	0	0	77	0	0	0
3.5	0	0	0	0	0	0	34	0	0	0
3.75	0	0	0	0	0	0	60	0	0	0
4	0	0	0	0	0	0	25	0	0	0
4.25	0	0	0	0	0	0	23	0	0	0
4.5	0	0	0	0	0	0	20	0	0	0
4.75	0	0	0	0	0	0	14	0	0	0
5	0	0	0	0	0	0	5	0	0	0
5.25	0	0	0	0	0	0	3	0	0	0
5.5	0	0	0	0	0	0	10	0	0	0
5.75	0	0	0	0	0	0	4	0	0	0
6	0	0	0	0	0	0	4	0	0	0
6.5	0	0	0	0	0	0	2	0	0	0
6.75	0	0	0	0	0	0	2	0	0	0
7	0	0	0	0	0	0	2	0	0	0

It is indeed striking that for a different sample of 20 000 observations such a similar result was obtained – even though in this case ten, not nine clusters, as before, were indicated by the silhouette method. As in the preceding example – three “bot” clusters were identified, two bigger ones featuring low degree of “botness”, and one – much higher. As before, several of the clusters were fully “pure” (“bot” clusters nos. 7 and 10, and “not-bot” clusters nos. 1, 2, 4 and 9). Of the “impure” clusters, the one with the highest degree of “impurity” (cluster no. 6) has merely some 0.12% of the ‘alien” observations.

The question, which arises in this context is: how does this translate into the predictive ability?

In order to answer this question yet another kind of analysis was performed, on the basis of the “pam” algorithm. Namely, Table 10. shows the (normalized) values of distances between the centroids of clusters, to be used in the classification exercise, based on the results from the clustering, performed with "pam". These clusters originate from three runs of the "pam" algorithm, the runs being denoted by capital letters A (clusters A1, A2,…), B (clusters B1, B2,…), and C (clusters C1, C2,…). For purposes of this classification exercise the nearest clusters were aggregated to form the "working clusters", whose acronyms appear in Table 10. as row and column labels (e.g. A2B1C4, meaning the aggregate of clusters A2, B1 and C4).

Table 10. Distances between the centroids of clusters, used in the classification exercise, resulting from clustering by the "pam" algorithm. The acronyms correspond to the output from aggregation of the closest clusters. Blue colour indicates "bot" clusters. The table is symmetric, hence only the upper half is shown

	A6	A2B1C4	A5B4C7	A7B8C3	A4B5C5	A9B6C6	A1B3C2	A3B2C1	A8B7C8
A6		0.0898	0.1298	0.0958	0.1193	0.3680	0.0560	0.0628	0.0457
A2B1C4		---	0.0219	0.0388	0.0456	0.2245	0.0106	0.0095	0.0464
A5B4C7			---	0.0509	0.0415	0.1138	0.0298	0.0367	0.0807
A7B8C3				---	0.0588	0.2174	0.0275	0.0406	0.0561
A4B5C5					----	0.1342	0.0377	0.0486	0.0744
A9B6C6						---	0.2203	0.2490	0.3022
A1B3C2							---	0,0024	0.0173
A3B2C1								---	0.0164
A8B7C8									---

First, let us emphasise that the distances, appearing in Table 10., differ by as much as two orders of magnitude (from the minimum – in this matrix – 0.0024 up to 0.368). *Then*, exactly in this context, it is extremely striking what is the distribution of the distances shown, namely, the absolute minimum of 0.0024 was calculated for a pair of clusters, labelled "bot" and "not-bot"! Not only this, but, in general, if we look at the cluster A1B3C2 then we see that it is quite, or even very close to most of the distinguished "not-bot" clusters! The distances between the "bot" clusters, even if truly small, are by no means clearly smaller than those to the "not-bot" clusters. *Third*, the "most distant" cluster of all (A9B6C6) is, again, highly surprisingly, a "not-bot" cluster, among as many as six "not-bot" clusters.

If we consider and interpret these results in the light of Figure 8., then they become much more understandable. The conclusion, though, is not very optimistic, and they can be subsumed in two statements: (A) there are some very distinct clusters, either "bot" or "not-bot", which can be relatively easily characterized as separate groups, and identified as such; (B) there are also "mixed" groups of data, in which distinction between "bots" and "not-bots" is very difficult – if at all possible. With respect to the latter, Figure 8. suggests that, in fact, the two kinds of agents may display behavior classified in a single group (or more than one group), "not-bots" forming the interior of the group, while "bots" appearing along their edges. This would be in agreement with what Table 10. tells us.

A confirmation of these conclusions is provided by the results from the predictive use of the set of clusters, roughly characterized in Table 10., which are rather poor. They amount, on the average to 85% of correct classifications, and even if we consider an extension of the strict indications so as to account for the second-best ones, we cannot achieve more than 92% of correctness.

Thus, a more refined approaches seem to be required in order to achieve truly better results.

6. Building classifiers - a case study

6.1. The description of the approach

The objective of the particular study here characterised was to deliver a two-tier approach to the analysis of human and bot activity data. At first, we partition the data into clusters. Next, we use a classifier to recognize elements in selected clusters. Classifiers are present only in these clusters, which contain mixed observations. The method is endowed with two parameters:

- k - the number of clusters to be extracted;
- t – the threshold level determining in which clusters we build a classifier;

The choice regarding the clustering and classification algorithms to be used in this approach can be made depending on the preferences of the model designers. We recommend the use of k-means as the clustering algorithm and decision tree as the classifier. The argument for choosing k-means is that it produces an output that is straightforward to interpret: clusters are represented with centroids, which inform about a typical instance falling into a given cluster. Decision trees are recommended for the same reason. In this case, the model is present in the form of a tree structure, where splits inform about the level of a feature that decides about class label assignment. The second tier – supervised clustering is treated as a fine-tuning step. Most of the heavy lifting is expected to be executed by the clustering algorithm.

An important factor when choosing the right algorithms is time complexity. The k-means algorithm is known to have a time complexity of $O(n^2)$, where n is the input data size (Pakhira, 2014), while the process of constructing decision tree using the popular CART algorithm can be estimated as $O(m*n*\log_2 n)$ where m is the number of attributes and n is the number of observations (Sani et al., 2018).

The decision, regarding the number of clusters to be created can be made in two ways. One is the application of a selected internal cluster validity index. The literature of the area offers a plethora of such cluster validity indexes, often evaluating the similarity of points belonging to the same cluster, dissimilarity of points categorised to different clusters, both these elements or other features related to the shape and characteristics of clusters. While reviewing particular cluster validity indexes, we often come across measures such as variance, entropy, distances, diameters, the density of clusters, and so on (Halkidi et al., 2001). Internal indexes measure clustered data itself. Internal cluster validity indexes promote clusters that are compact and regular: well-separated clouds with a small variance between members of the cluster. Internal cluster validity indexes are usually based on similarity measures corresponding to similarity measures used in clustering algorithms (Kryszczuk and Hurley, 2010).

The second criterion is to apply an external cluster validity criterion, that is, to take the benefit of the labels "bot"/"human" and choose the partitioning that assures the best separation of "bots" and "humans" from the beginning.

The former approach (internal index) is more universal, however, in practice we obtain quite different results depending on the choice of the index and the preprocessing scheme of the data set. In our experiments with internal indexes, we often obtained recommendations concerning 2, 3, and 6 clusters for the analyzed data set. Let us look at the examples of partitioning obtained for these three cases on a randomly selected balanced sample of instances, shown in Table 11. below.

Table 11. Three examples of clustering results for a balanced sample of "bots" and "humans" as the preliminary step in the cluster-then-classify procedure

a) Split into k = 2 clusters

cluster id	cardinality	bot share (%)
C1	21749	49.91
C2	4927	50.40

b) Split into k = 3 clusters

cluster id	cardinality	bot share (%)
C1	7089	78.25
C2	3115	39.78
C3	16472	39.78

c) Split into k = 6 clusters

cluster id	cardinality	bot share (%)
C1	4975	90.43
C2	1968	60.57
C3	13637	33.17
C4	1043	37.20
C5	4177	44.51
C6	876	100.00

The average separation for these three cases is equal to 50.25%, 66.23%, and 72.69%, respectively. Thus, we choose the split into six clusters as the most suitable for this data set. In this case, we obtain one cluster composed of bots only (C6), one cluster almost exclusively made of bots (C1), and four rather mixed clusters. Looking at the obtained clusters, we can set the threshold level to 0.1, meaning, in this case, that decision trees will be built for four clusters (C2, C3, C4 and C5). An alternative way of determining the threshold level would be to set it to the desired recognition rate.

To evaluate the quality of bot/human identification, we need to choose appropriate measures. Since the data is skewed, besides standard accuracy measure, we propose to apply the F1 score, which is the harmonic mean of precision and recall. The formula for accuracy is given as:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where TP is the number of correctly identified bots, TN is the number of correctly identified humans. FP and FN are incorrectly identified humans and bots, respectively. Based on the values of TP, TN, FP, and FN we can compute precision and recall which are:

$$precision = \frac{TP}{TP+FP} \tag{2}$$

$$recall = \frac{TP}{TP+FN}. \tag{3}$$

And finally, F1 is given as:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision+recall}. \tag{4}$$

The chosen quality measures allow for verifying if a recognition mechanism is not leaning towards one of the two classes, which may happen for skewed data.

We propose to execute the experiments in the following manner.

- The training set should be made of randomly selected samples from one day. There should be an equal number of bots and humans in this sample (it should be balanced).
- Test set, same day: bots and humans from the same day as the training set, but samples disjoint from the train set. Balanced set.
- Test set, next day: bots and humans from the next day (train set is from the previous day). This set is not balanced.

6.2. The exemplary results

Using the above-mentioned sets allows for estimating the true quality of the procedure. Results, concerning the train set will not be shown, since the train sets were used to construct the models. Results concerning test sets of the two kinds are presented in Table 12. The constructed models were all based on six clusters.

Table 12. Accuracy and F1 measure (in %) for models constructed using samples from different days. All models are built with k = 6 and t = 0.9

	clusters				clusters + decision trees			
	test, same day		test, next day		test, same day		test, next day	
train date*	F1, %	accuracy, %	F1, %	accuracy, %	F1, %	accuracy, %	F1, %	accuracy, %
3	86.27	85.21	66.64	83.12	86.27	85.21	66.64	83.12
4	61.17	68.99	54.72	87.93	87.50	87.34	55.34	84.54
5	90.52	90.05	59.16	83.66	90.52	90.05	59.16	83.66
6	75.60	77.94	55.55	86.51	93.68	93.53	78.36	93.36
7	81.43	83.36	67.61	90.96	88.10	88.68	75.33	92.54
8	87.42	87.62	65.77	87.78	95.00	95.02	85.43	95.43
9	89.01	89.31	66.50	88.02	93.51	93.41	72.44	89.65
10	63.90	70.63	51.19	81.63	83.64	85.05	71.58	89.36
11	62.48	69.41	51.57	81.91	80.33	82.46	68.44	88.23
12	50.00	64.08	41.58	82.84	79.18	79.60	59.63	81.40
13	64.56	71.07	53.40	84.30	81.80	82.64	65.92	86.63
14	67.00	72.42	49.75	82.17	82.91	83.51	66.99	87.08
15	68.20	72.37	50.13	80.75	78.92	80.51	61.40	85.02
16	53.98	65.84	44.11	83.77	80.13	81.81	64.50	86.91
17	60.45	67.20	45.85	80.78	80.87	80.86	57.84	81.72
18	65.46	71.64	50.04	83.35	80.26	81.19	60.39	84.68

* August 2020

Results, provided in Table 12., show, first, that there is concept drift in the data. The next day test set turned out to be altogether much more challenging than the same day test data. This difference can be up to 10%.

We observe that the quality of the model varies from one day to another. In the worst-case scenario, that is, when the model was trained on data coming from August 12th, 2020, we obtained the accuracy of 64.08% for the same day test set and 82.84% for the next day test set when using only clusters.

Adding decision trees to the cluster-based model in the vast majority of cases improved the recognition rate. On average, accuracy on the same day test increased from 75.45% to 85.68%, F1 on the same day test increased from 70.47% to 85.16%. On average, accuracy on the next day test increased from 84.34% to 87.08%, F1 on the same day test increased from 54.60% to 66.84%. Thus, the improvement over pure clustering is indeed very significant. Achieved recognition rates are satisfactory knowing that the data is drifting.

The use of both accuracy and the F1 score is necessary since we observe that the values of accuracy are higher than those of the F1 score and we need to monitor this difference.

7. Summary and conclusions

In this paper, we have presented an outline of the problem of distinguishing "natural" and "artificial" traffic on the web, related to the advertising content. Besides the presentation of the problem and its peculiarities, we illustrated the steps in the respective analytic approach, with indication of the potential choices on each of these steps.

We take the benefit of the features describing the actor's behaviour. The observations that are analysed are labelled with a rule-based expert-developed tool. In the first stage of work, the variables are defined, which might be used in further analysis. In particular, an attempt was made to determine the most effective set of variables, to be then used in the solution of the problem. A number of illustrations was provided, concerning this analysis, related to the temporal behaviour of the variables, their interrelations, as well as principal component analysis.

Then, the results from the clustering approach were presented. Of special interest was the conclusion, already suggested in the first stage of the study, that *many more groups of behaviour patterns ought to be distinguished than just two, i.e. bots and humans*. Moreover, the second essential conclusion, also associated with some of the results from the first stage of investigations could be formulated, valid for virtually all of the analysed samples, namely that while *some of these behaviour pattern groups are relatively "pure", i.e. containing virtually solely the observations of one kind (i.e. either bots or humans), some important groups contained both of them within quite homogeneous subsets of observations*. Both these conclusions led to further investigations, conducted according to a different methodology.

The approach applied in the subsequent stage was a combination of unsupervised and supervised learning. At first, we partitioned the data into clusters. Next, in selected clusters, we built decision trees. The number of clusters and the decision when to build a decision tree within a cluster are the parameters of the proposed model. The former can be determined using an internal cluster validity index, for instance, the silhouette index. The latter can be determined as a target accuracy one wishes to obtain. The model can be instantiated with the use of k-means and CART decision tree which both have the advantage of being straightforward to interpret by a human being. Because the data is skewed, it is necessary to trace both the accuracy and the F1 (or a similar measure or quality) of the constructed models. We illustrated the proposed procedure on a selected real-world data set. The objective was to obtain satisfying results knowing that the data is subjected to drifts. The achieved average accuracy reaches about 87% on a test set originating from a date different than the training set date.

References

Aberathne I. and Walgampaya C.: Smart mobile bot detection through behavioral analysis. *Advances in Data and Information Sciences*. Springer, 2018, 241-252.

Cai Y., Yee G. O. M., Gu Y. X. and Lung C.-H.: Threats to online advertising and countermeasures: A technical survey. *Digital Threats: Research and Practice* 1(2), 2020. Available: https://doi.org/10.1145/3374136

Ester M., Kriegel H.-P., Sander J. and Xu X.-w.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: E. Simoudis, J.-w. Han, U. M. Fayyad (eds.) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 1996, 226-231.

Gagolewski M., Bartoszuk M. and Cena A. Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences* **363**, 2016, 8-23.

Gajewski M., Hryniewicz O., Jastrzębska A., Opara K., Owsiński J. W., Zadrożny S., Kozakiewicz M. and Zwierzchowski T.: Explainable identification of bots from web activity logs, 2021 (submitted).

Gajewski M., Hryniewicz O., Jastrzębska A., Kozakiewicz M., Opara K., Owsiński J. W., Zadrożny Sł. and Zwierzchowski T.: Assessing the Share of the Artificial Ad-Related Traffic: Some General Observations. Chapter 26 in: C. Ciurea et al. (Eds.) *Education, Research and Business Technologies. Smart Innovation, Systems and Technologies* **276***,* Springer Nature Singapore Pte Ltd., 2022.

Halkidi M., Batistakis Y. and Vazirgiannis M.: On clustering validation techniques. *Journal of Intelligent Information Systems* **171**(2-3), 2001, 107-145.

Khattak S., Ramay N. R., Khan K. R., Syed A. A. and Khayam S. A.: A taxonomy of botnet behavior, detection, and defense. *IEEE Communications Surveys & Tutorials* **16**(2), 2014, 898-924.

Kryszczuk K. and Hurley P.: Estimation of the number of clusters using multiple clustering validity indices. In: *Multiple Classifier Systems*. *Lecture Notes in Computer Science* **5997**, Springer: Cham, 2010, 114-123.

Ling R. F.: On the theory and construction of k-clusters. *The Computer Journal* **15**(4), 1972, 326-332. doi:10.1093/comjnl/15.4.326

Mouawi R., Elhajj I. H., Chehab A. and Kayssi A.: Crowdsourcing for click fraud detection. *EURASIP Journal on Information Security* 11, 2019, https://doi.org/10.1186/s13635-019-0095-1

Owsiński J. W., Kacprzyk J., Opara K. R., Stańczak J., Zadrożny S.: Using a Reverse Engineering Type Paradigm in Clustering. An Evolutionary Programming Based Approach. *Fuzzy Sets, Rough Sets, Multisets and Clustering,* 2017, 137-155.

Owsiński J. W., Stańczak J., Opara K., Zadrożny S. and Kacprzyk J.: *Reverse Clustering. Formulation, Interpretation and Case Studies*. *Studies in Computation Intelligence* **957**, Springer International Publishing, 2021

Pakhira M. K.: A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting, *2014 International Conference on Computational Intelligence and Communication Networks*, Bhopal, India, 2014, 1047-1051, doi: 10.1109/CICN.2014.220.

Sani H. M., Lei C. and Neagu D.: Computational complexity analysis of decision tree algorithms. In: M. Bramer and M. Petridis (eds) *Artificial Intelligence XXXV. SGAI 2018. Lecture Notes in Computer Science*. Springer: Cham, **11311**, 191-197.

Thejas G. S., Dheeshjith S., Iyengar S. S., Sunitha N. R. and Badrinath P.: A hybrid and effective learning approach for Click Fraud detection. Machine Learning with Applications 3, 2021, https://doi.org/10.1016/j.mlwa.2020.100016

Websites

http://multirbl.valli.org

https://mxtoolbox.com/

DISTINGUISHING THE ARTIFICIAL AND THE GENUINE AD-RELATED TRAFFIC: MAIN OBSERVATIONS AND EXEMPLARY RESULTS

Abstract: We describe the essential aspects of the project, aimed at developing a methodology for distinguishing the artificial, i.e. automatically generated, internet traffic, from the genuine ones, i.e. produced by humans, regarding the advertising on the web. So, we first present the nature of the problem, including its rough business rationality and then the key characteristics of the relevant internet traffic. This is followed by the outline of the set of methodologies used on the project, and then an excerpt of the results is presented with the corresponding comments and discussion. Finally, the conclusions, technical and of a more general character, are forwarded.

Keywords: internet, advertising, artificial traffic, bots, classification, clustering, data analysis

ODRÓŻNIANIE SZTUCZNEGO I NATURALNEGO RUCHU INTERNETOWEGO DOTYCZĄCEGO REKLAM: ZASADNICZE USTALENIA I PRZYKŁADOWE WYNIKI

Streszczenie: Opisujemy w artykule zasadnicze aspekty projektu, którego celem było opracowanie metodyki rozróżniania sztucznego, tj. generowanego automatycznie, ruchu internetowego od rzeczywistego, czyli będącego wynikiem działania autentycznych użytkowników internetu, w kontekście pokazywanych w internecie reklam. Najpierw pokazujemy, na czym polega zagadnienie i jego aspekt komercyjny, a następnie główne cechy odnośnego ruchu internetowego. Następnie zarysowujemy metodyki, wykorzystywane w projekcie oraz przykłady otrzymywanych przy ich pomocy wyników, wraz z towarzyszącymi komentarzami i dyskusją. Na koniec przedstawiamy wnioski i rekomendacje z przeprowadzonych badań.

Słowa kluczowe: internet, reklama, sztuczny ruch, boty, klasyfikacja, analiza danych