big data and cognitive computing

MDPI

Article

Experimental Evaluation: Can Humans Recognise Social Media Bots?

Maxim Kolomeets [1,2], Olga Tushkanova [2], Vasily Desnitsky [2,†], Lidia Vitkova [2,†] and Andrey Chechulin [2,*,†]

[1] School of Computing, Newcastle University, Newcastle upon Tyne NE4 5TG, UK; maksim.kalameyets@newcastle.ac.uk or kolomeec@comsec.spb.ru

[2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, 14th Line of V.O. 39, St. Petersburg 199178, Russia; tushkanova@comsec.spb.ru (O.T.); desnitsky@comsec.spb.ru (V.D.); vitkova@comsec.spb.ru (L.V.)

* Correspondence: chechulin@comsec.spb.ru

† These authors contributed equally to this work.

Abstract: This paper aims to test the hypothesis that the quality of social media bot detection systems based on supervised machine learning may not be as accurate as researchers claim, given that bots have become increasingly sophisticated, making it difficult for human annotators to detect them better than random selection. As a result, obtaining a ground-truth dataset with human annotation is not possible, which leads to supervised machine-learning models inheriting annotation errors. To test this hypothesis, we conducted an experiment where humans were tasked with recognizing malicious bots on the VKontakte social network. We then compared the "human" answers with the "ground-truth" bot labels ('a bot'/'not a bot'). Based on the experiment, we evaluated the bot detection efficiency of annotators in three scenarios typical for cybersecurity but differing in their detection difficulty as follows: (1) detection among random accounts, (2) detection among accounts of a social network 'community', and (3) detection among verified accounts. The study showed that humans could only detect simple bots in all three scenarios but could not detect more sophisticated ones (p-value = 0.05). The study also evaluates the limits of hypothetical and existing bot detection systems that leverage non-expert-labelled datasets as follows: the balanced accuracy of such systems can drop to 0.5 and lower, depending on bot complexity and detection scenario. The paper also describes the experiment design, collected datasets, statistical evaluation, and machine learning accuracy measures applied to support the results. In the discussion, we raise the question of using human labelling in bot detection systems and its potential cybersecurity issues. We also provide open access to the datasets used, experiment results, and software code for evaluating statistical and machine learning accuracy metrics used in this paper on GitHub.

Keywords: social bot detection; bot evolution; disinformation; ground-truth problem; cybersecurity

check for updates

Citation: Kolomeets, M.; Tushkanova, O.; Desnitsky, V.; Vitkova, L.; Chechulin, A. Experimental Evaluation: Can Humans Recognise Social Media Bots? *Big Data Cogn. Comput.* **2024**, *8*, 24. https://doi.org/10.3390/bdcc8030024

Academic Editors: Min Chen and Michele Melchiori

Received: 7 December 2023
Revised: 16 February 2024
Accepted: 20 February 2024
Published: 26 February 2024

1. Introduction

Social media bots are widely used for disinformation, fraud, reputation cheating, blackmail, hacking of recommendation systems, and other malicious activity in social media. For that reason, bot combatting has become a highly relevant topic in cybersecurity and social sciences, where researchers are trying to create novel combat techniques and estimate bot impact on social processes. With the growing role of machine learning in bot detection and social network analysis, we want to discuss the "ground-truth label problem" that is related to the "hidden inefficiency" of some supervised bot detection systems due to the low quality of training datasets.

We define the "ground-truth label problem" as follows—*the quality of social media bot detection systems based on supervised machine learning and human annotators may not be as accurate as researchers claim because humans' ability to detect bots is similar to random guess and detection models inherit their errors*.

The critical point of the "ground-truth label problem" is that researchers suppose that humans can detect bots. Therefore, there are many papers where bot detection systems were trained and tested as supervised machine learning models using datasets labelled by human annotators and where human answers were considered near "ground-truth" data. In this paper, we question the ability of a human to recognise a bot or at least several types of bots. We consider this problem from several points of view, including (1) the experimental evaluation of human bot detection ability over different bot detection scenarios and bot types, (2) the theoretical evaluation of bot detection systems that are trained on datasets labelled by human annotators, (3) and the consequences of such bot detection systems usage.

The paper consists of the following sections: In Section 2, we analyse the existing labelling methods and place human annotation techniques among them. In Section 3, we present the research approach, which includes the following: (Section 3.1) testing of several hypotheses about the human ability to recognise bots, (Section 3.2) an approach for the evaluation of a classifier trained on a dataset labelled by humans, (Section 3.3) a description of the datasets that we used in the experiment, (Section 3.4) an experiment design description. In Section 4, we present the results of the experiments. In Section 5, we discuss the paper's results and limitations and make some assumptions.

2. Related Work

The evolution of bot technology has made it increasingly difficult for even the most discerning users to identify them. Modern bots exhibit sophisticated behaviors, including posting content, interacting with users, and participating in discussions in a manner that closely resembles human activity [1,2].

A general lack of awareness and education regarding the presence and capabilities of bots further complicates efforts to identify them. Many users are simply unaware of the extent to which bots can mimic human behaviors or the signs that may indicate an account is not operated by a human [3].

At the same time, social bot detection is a well-established area of social network analysis. As social media plays an increasingly important role in society, the identification of bots and the distinction between natural and distorted social processes are becoming essential for social media security and online safety. Bots can affect the ratings of products in online markets, influence stock exchange prices, participate in political propaganda, and engage in malicious activities such as fraud. As a result, many researchers are attempting to develop bot detection solutions to analyse social media accounts and determine whether or not they are bots to take countermeasures that can block illegitimate accounts or reduce the speed and extent of information dissemination by bots.

To be more precise, it is essential to note various bot definitions, as researchers from different areas mention social bots differently [4]. Therefore, when we refer to bots in this study, we do not only mean automated accounts controlled by software but any account that someone can purchase to carry out controlled malicious activities, including hacked accounts, fake identities that are controlled by human operators, genuine-users who are paid to perform malicious actions, and any other types of bots that may emerge on the bot market [5]. This broad definition captures the full range of potential threats posed by malicious actors, regardless of the specific means by which they are carried out.

One of the most challenging aspects of detecting bots is the phenomenon of bot evolution [4,6]. Many researchers have pointed out that bots constantly evolve, making it increasingly difficult to identify each new wave of social bot accounts. This creates a scenario resembling an arms race [7], where bot detection and bot creation techniques are constantly competing to stay ahead of each other. The evolution of bots can occur in various ways, including changes in their behavioural patterns, new tactics and strategies, and advancements in their programming and infrastructure. As a result, bot detection methods must continually adapt and evolve to keep up with the ever-changing landscape of social media bot activity and types.

The detection and combatting of social bots have seen significant advancements through the development of adaptive bot detection methods. Recent research has introduced a Conditional Adversarial Learning Framework, CALEB, utilizing conditional generative adversarial network (CGAN) to simulate bot evolution. This approach enhances the detection of new bot types by generating realistic synthetic instances of various bot types, offering a performance boost in detecting unseen bots [1]. Furthermore, the BotMatch methodology utilizes a semi-supervised recursive nearest neighbors search to map emerging social cybersecurity threats, providing a novel approach to similarity-based bot detection [8]. An interesting contribution introduced a community-level bot detection framework, BotPercent, which estimates the percentage of bot accounts within specific online communities. This approach marked a notable evolution from focusing on individual bot identification towards assessing bot influence across broader social groups [9].

The challenge of fighting evolving Twitter spammers has been empirically evaluated, leading to the design of new detection features that significantly improve the detection rate and reduce false positives [10,11]. Machine learning has also been applied for profiling social network users to detect bots, demonstrating the potential of automated user profiling for detecting fake accounts with high accuracy [12]. Another study presented an algorithm for classifying Twitter accounts as bots or non-bots, emphasizing the selection of the most efficient algorithm based on detection accuracy [13].

However, a fundamental problem with these “human-annotation based tools” is the lack of ground-truth bot labels in training samples. Some researchers also mention these concerns [14,15]—the limiting factor in advancing bot detection research is the lack of availability of robust, high-quality data, caused by “simplistic data collection and labeling practices”. To highlight the prevalence of this problem, we aggregated several bot labelling techniques and their popularity (see Figure 1) from the review of papers presented in [16]. These techniques include the following:

1. Manual labelling: datasets that are labelled by human annotators that are trying to recognise bots [17–19];
2. Purchasing: datasets obtained directly from bot-traders/owners [20–22] or by honeynets [23], when researchers ask bot owners to share bot identifiers or purchase such bots to find it;
3. Anomaly behaviour: datasets labelled on the basis of abnormal metrics [24] (such as exceeding spam ratio thresholds [10]);
4. Declaration: datasets where labelling is performed by a third party (such as suspended accounts by social network defence systems [25], open-access datasets without mention of a labelling method, etc.);
5. Other: datasets where researchers used their own labelling method or did not mention which method they used (these were not considered further).

One of the most popular methods for bot labelling is manual labelling, which heavily relies on the ability of the annotator to differentiate between bots and genuine-users. The declaration method also depends on the classifier’s effectiveness in dataset formation. The suspicious behaviour method may miss sophisticated bots with complex behaviour patterns, particularly new waves of bots that may use AI techniques to appear more natural. Among these methods, the purchasing method is the closest to the ground-truth as researchers buy bots directly from bot owners.

In this paper, we focus on analysing the efficiency of manual labelling, which has become less effective due to the evolution of social bots, when bots become so sophisticated that the ability of humans to detect a bot accurately is questionable. This is supported by bot-traders separating bots [5] into different categories according to their quality, assuming that bot detection systems cannot recognise high-quality bots operated by humans or partly generated by AI. Therefore, AI bot detection tools may inherit the errors of the annotators, which may lead to false negatives and false positives. These errors can lead to discrimination, false accusations, and loss of trust in communities when the detection tool recognises a bot attack where there is natural activity of genuine-users [26].

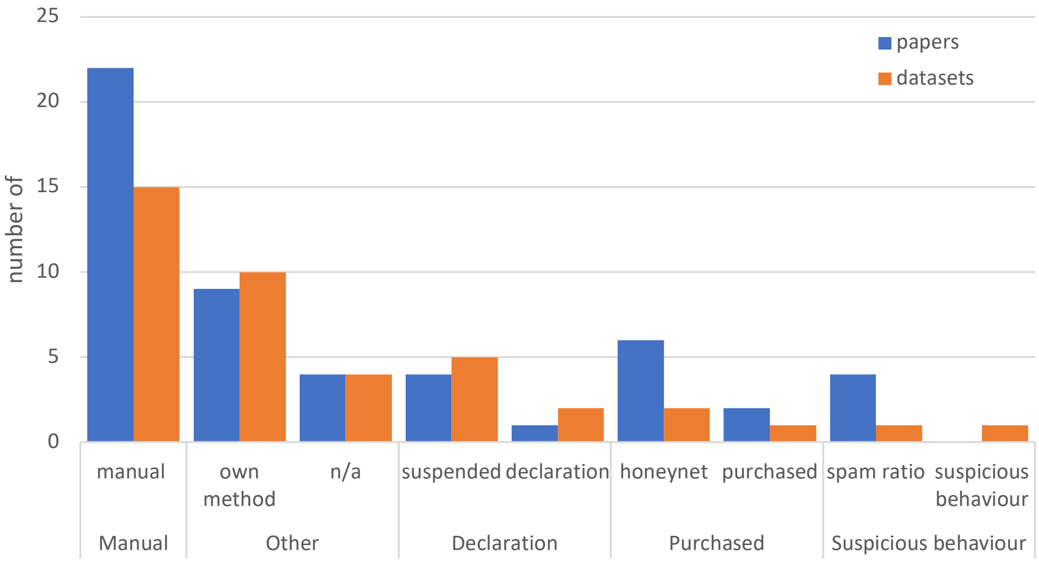


Figure 1. Distribution of labelling methods in 41 papers reviewed in [16]—note that most of the researchers use datasets where bots were labeled by humans.

Therefore, we propose an experimental evaluation to investigate and measure how humans can recognise social bots of different types and how it affects bot detection systems that leverage supervised machine learning techniques.

3. Materials and Methods

To measure a human bot detection ability and how well one can label bots with human annotation technique, we propose estimating the difference between human answers and ground-truth labels as follows: we conduct several experiments comparing the *ground-truth bot labels* obtained with the purchased method with *bot labels obtained by manual annotation* by humans.

3.1. Hypotheses about the Human Ability to Detect Bots

The problem of detecting bots is a binary classification problem, with the aim of determining whether a given account belongs to a bot (1) or not (0). The success of the human ability to detect bots can be evaluated by comparing it to that of a binary random classifier, where success is defined as correctly identifying a bot.

To evaluate this, we establish null and alternative hypotheses as follows:

Hypothesis 1: *Humans can detect bots.*

***H_0*:** The probability of humans detecting a bot is not better than that of a random classifier. The success rate for humans is ≤ 0.5.

***H_1*:** The probability of humans detecting a bot is better than that of a random classifier. The success rate for humans is 0.5.

If: p-value $< 0.05 \longrightarrow$ reject H_0.

To consider the variability of the different types and qualities of sophisticated bots, we must address the question of whether humans can distinguish bots of specific qualities. It is reasonable to assume that human recognition abilities may differ depending on the quality of the bot in question.

Therefore, we formulate several alternative hypotheses based on the type of bot quality:

Hypothesis 2: *Humans can detect bots of specific quality "X".*

$\boldsymbol{H_0}^{\boldsymbol{X}}$**:** The probability of humans detecting a bot with quality "X" is not better than that of a random classifier. The success rate for humans is $\leq$0.5.

$\boldsymbol{H_1}^{\boldsymbol{X}}$**:** The probability of humans detecting a bot with quality "X" is better than that of a random classifier. The success rate for humans is >0.5.

If: p-value $< 0.05 \longrightarrow$ reject H_0

Moreover, the scenario in which bot detection takes place can also affect the ability of humans to detect bots. For example, humans may perform clustering instead of classification when trying to recognise bots, comparing given accounts with each other and separating them into bots and genuine-users. For example, it may be easier to detect bots among profiles where genuine-users' accounts have a long history, many photos, and look natural. On the opposite hand, annotators can make more errors when genuine-users' accounts look suspicious—when they are anonymised, relatively new, or lacking content. To test it, we propose that the human ability to allocate bots depends on the type of genuine-user accounts presented in the sample. We will examine the following three distinct scenarios:

1. Random account analysis: an annotator identifies bots from a representative sample of social network accounts. In this situation, humans detect bots from all social network accounts. Since only a small percentage of social network accounts are active, most accounts in the representative sample may have empty profiles and/or no activity. Thus, annotators may mistakenly categorize some bots as more similar to genuine-users during subconscious clustering;
2. Shifted account analysis: an annotator detects bots among a shifted sample of accounts collected from the same community. In this scenario, humans detect bots from some group of individuals with homophily. As these accounts have a shift in their features (users have similar characteristics, such as belonging to the same location or age group, or having other similar features due to homophily), it may be easier for an annotator to cluster a dataset into bots and not bots parts.
3. Verified users analysis: an annotator detects bots among accounts of people known to a researcher. In this case, humans detect bots from a group of individuals with very strong homophily. As these accounts will have strong homophily (all these people are acquaintances of the researcher) and well-filled profiles, it can be easier for annotators to cluster the dataset and make fewer mistakes.

These scenarios also reflect the dataset collection strategies (how a researcher collects a sample of genuine-users). To test the scenario effect, we propose the following hypothesis:

Hypothesis 3: *Humans can detect bots in specific scenario "Y".*

$\boldsymbol{H_0}^{\boldsymbol{Y}}$**:** The ability of humans to recognise bots in scenario "Y" is no better than that of a random classifier. The probability of success for humans is $\leq$0.5.

$\boldsymbol{H_1}^{\boldsymbol{Y}}$**:** The ability of humans to recognise bots in scenario "Y" is better than that of a random classifier. The probability of success for humans is >0.5.

If: p-value $< 0.05 \longrightarrow$ reject H_0.

Since each hypothesis analyses the number of successful bot detection events, which form a binomial distribution, we can test these hypotheses with a binomial test. If the p-value of the test is <0.05, we reject the null hypothesis and conclude that humans can recognise bots. We can also construct confidence intervals for each test and visualise them to see how much the bot detection ability differs for different bot types and detection scenarios.

3.2. Hypothetical Classifier Efficiency Metrics

In addition to assessing the human ability to detect bots, we propose an evaluation of the impact of annotator errors on the effectiveness of a hypothetical bot detection classifier trained on human-annotated datasets.

The effectiveness of a real classifier depends on the following two factors: (1) the model's efficiency in separating accounts labelled as bots from those labelled as genuine-users, and (2) the accuracy of the dataset labels.

To conduct a theoretical experiment, we propose considering a hypothetical classifier with an accuracy of 1. In this hypothetical case, the real accuracy is limited only by the accuracy of the data labelling. We can estimate the accuracy of the dataset labelling using basic machine learning metrics that are robust to bot–user balance:

1. True positive rate (TPR): expresses how many bots were recognised correctly and equation $\frac{TP}{TP+FN}$.
2. True negative rate (TNR): expresses how many genuine-users were recognised correctly and equation $\frac{TN}{TN+FP}$.
3. Balanced accuracy (BACC): an integrated measure that considers both TPR and TNR and equation $\frac{TPR+TNR}{2}$.

A hypothetical perfect classifier that achieves an accuracy of 100% when trained and tested on human-annotated labels with errors would still have an efficiency metric no better than our estimations based on the ground-truth labels. In other words, the accuracy of the data labelling fundamentally limits the potential efficiency of the classifier. Therefore, we can estimate the efficiency of a real classifier by multiplying the efficiency metric of the machine learning model by our estimations of the efficiency metric based on the ground-truth labels:

$$real_m = model_m \times hypothetical_m, \tag{1}$$

where $hypothetical_m$ represents the efficiency of a hypothetical classifier with estimated $BACC = 1$ while the real accuracy equation accuracy of dataset labelling and $model_m$ represents the efficiency of the machine learning model.

3.3. Datasets Used

To verify our hypotheses, we required a dataset containing ground-truth bot labels and genuine-user labels. We created MKVKTT21 dataset, based on the MKVK2021 dataset from the VKontakte social network, which was previously obtained and described in our previous research [21]—they both are publicly available on GitHub [27]. VK ranks [28] 8th in worldwide popularity as of March 2023 and is the number 1 social network in Russia, with a broad coverage of 84% of the Russian-speaking internet segment [29]. To collect the bot identifiers, we identified three bot-trader companies that offer promotion services on social media and provide bots of varying quality. To gather the bot labels, we posed as customers and created three fake groups on VKontakte. We then purchased likes ("thumbs up") for our posts from each bot-trader company sequentially. After each task was completed, we (1) collected the identifiers of the accounts that gave a like, (2) deleted the posts, (3) created new ones, and (4) purchased likes with another quality of bots or from another bot-trader company. Using this procedure, we obtained ground-truth bot labels, with bot identifiers categorised by bot-trader company and bot quality. The quality of bots was aggregated in ascending order of declared quality and price as follows: LOW, MID, HIGH, and LIVE; where LOW quality bots are the cheapest bots available, and LIVE quality bots are presumably operated by humans. The summary of collected bots is presented in Table 1, where we indicate the bot-trader/company code, the number of bots in each sample, and the quality label declared by the company.

Companies PARIK and OLENI are "bot activity" shops, where customers make a task, and bot-trader performs requested actions in less than an hour or even 1 min. These companies define bot quality [5] as follows:

1. LOW quality—software-created bots with a minimum profile fill;
2. MID quality—humans or software-created bots with a filled profile;
3. HIGH quality—account of genuine-users that became bots (by account stealing or buying).

Company MARSHRUTKA is a "bot activity" stock exchange where a buyer publicly gives a task, and anybody with an account that meets conditions (age, location, etc.) can perform necessary actions for money. Such bots were labelled as LIVE quality (note that in MKVKTT2021, in comparison to the original MKVK2021 dataset [27], for MARSHRUTKA company, we replace labels with LIVE, as it is a bot exchange [5] platform).

Table 1. Bot dataset with ground-truth labels and bot quality types that was used in the experiment.

Company Code	Quality	Count
PARIK	LOW	295
	MID	301
	HIGH	298
OLENI	LOW	303
	MID	301
	HIGH	304
MARSHRUTKA	LIVE	302
	LIVE	357

As genuine-users, we used identifiers of different sets of VKontakte accounts described in Table 2. It includes "Groups shifted" genuine-users from MKVK2021 dataset, and "Random", "Student" genuine-users, which we have additionally added.

Table 2. User dataset with sample properties that was used in the experiment.

Sample Name	Sample Type	Empty Profiles Number	Homophily	Scenario	Count
RANDOM	Representative	HIGH	No	Find bot in a whole social network	100,000
GROUPS_SHIFTED	Shifted	MID	Yes	Find bot in large community	2460
STUDENTS	Shifted	LOW	Yes, strong	Find bot in a small community	1077

The "Random" users in our dataset were selected as a representative sample of 100,000 identifiers from the general population of VKontakte, which includes 600 million accounts (at the time of the experiment). However, some of these accounts may be inactive or even bots, as we cannot determine the actual ratio of bots to genuine-users in VKontakte's general population.

On the other hand, the "Groups shifted" users are a subset of 2460 users from our previous study [21], where we collected users who performed "like" activity on ten groups [27]. These users have more active accounts and fewer empty ones, but there may still be bots since some groups may buy bots for content promotion. Furthermore, this sample does not represent the general population, as it only represents users with specific interests.

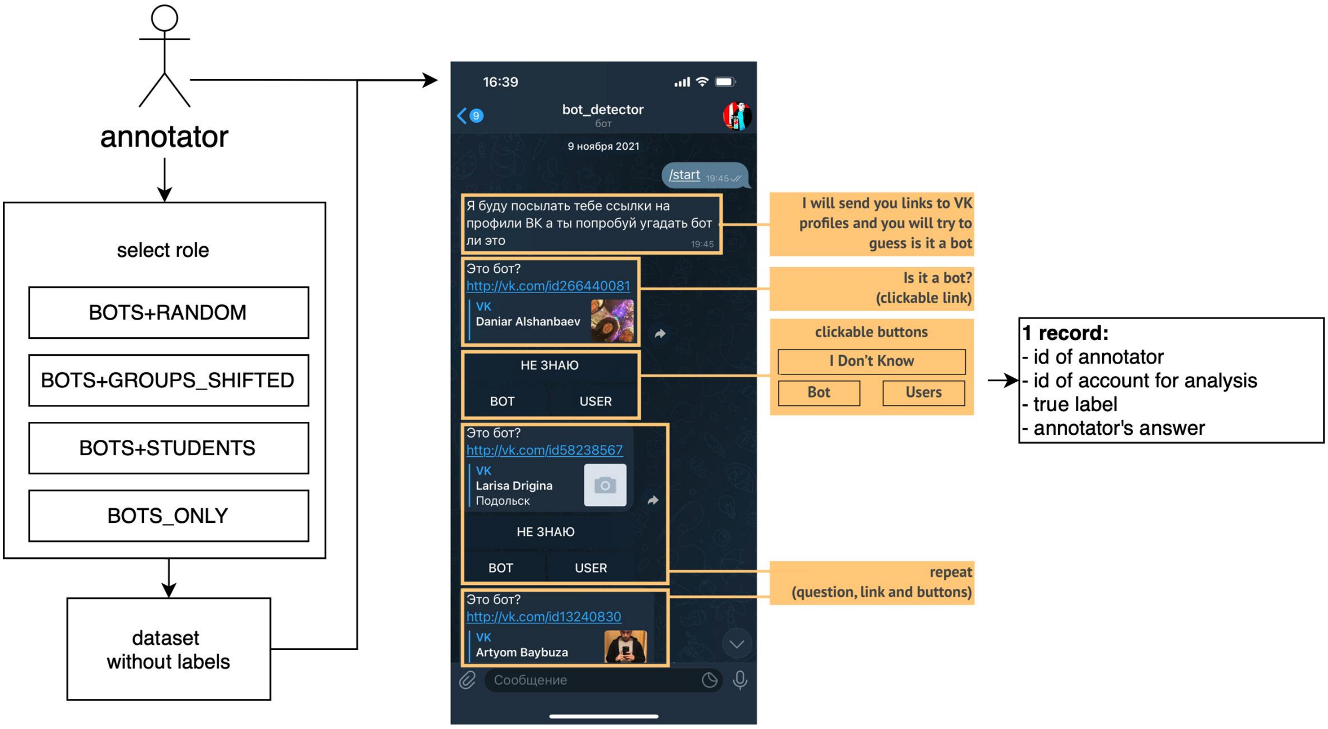
The "Student" users were collected from the study [30] where researchers collected VKontakte profiles of their students who provided links to their accounts for experimentation purposes. This set has the advantage of having "not a bot" labels that are ground-truth, as each account belongs to a verified student as determined by a university professor. However, this set is also highly unrepresentative as it only includes people of one age group with very similar interests.

In summary, the user labels in our collected dataset cannot be considered as ground-truth as we cannot be certain that there are no bots among them, except for the "Student" set. Therefore, we do not use these accounts in hypothesis testing but in evaluating hypothetical classifiers.

3.4. Experiment Design

To test hypotheses, we conducted an experiment where students manually labelled accounts of the VKontakte social network. For that, we developed a "bot_detector" annotation tool that works in Telegram messenger.

"Bot_detector" is a conversation bot that provides a user-friendly interface for account labelling that logic is presented in Figure 2.



16:39

bot_detector
бот

9 ноября 2021

/start 19:45

Я буду посылать тебе ссылки на профили ВК а ты попробуй угадать бот ли это 19:45

Это бот?
http://vk.com/id266440081
VK
Daniar Alshanbaev

НЕ ЗНАЮ
BOT USER

Это бот?
http://vk.com/id58238567
VK
Larisa Drigina
Подольск

НЕ ЗНАЮ
BOT USER

Это бот?
http://vk.com/id13240830
VK
Artyom Baybuza

Сообщение

Figure 2. Experiment design schema—we separated annotators into roles that are different in bot detection difficulty. Yellow boxes represent the English translation of the interface.

When the annotator starts to chat with "bot_detector", it provides a link to the VKontakte account and three buttons "I do not know", "Bot", and "User". Annotators must open the link, view the account, and press the selected button after deciding. After the annotator presses the button, "bot_detector" repeats the procedure—provides a link to another account with the same three buttons.

On the server side, when a user starts the conversation, we select one of the following roles for the user:

1. BOTS+RANDOM—provide to annotator only bots and users from the representative sample;
2. BOTS+GROUPS_SHIFTED—provide to annotator only bots and users from the shifted sample;
3. BOTS+STUDENTS—provide to annotator only bots and users from the sample of students;
4. BOTS_ONLY—provide only bots to annotator.

After that, for each annotator, we randomly form a sample with 50% of bots and 50% of users (except BOTS_ONLY role, where 100% are bots). The *Bots subsample* includes an equal number of bots of all types (see Table 1). The *Users subsample* has user accounts according to the role; therefore, we split annotators into different roles that correspond to detection scenarios.

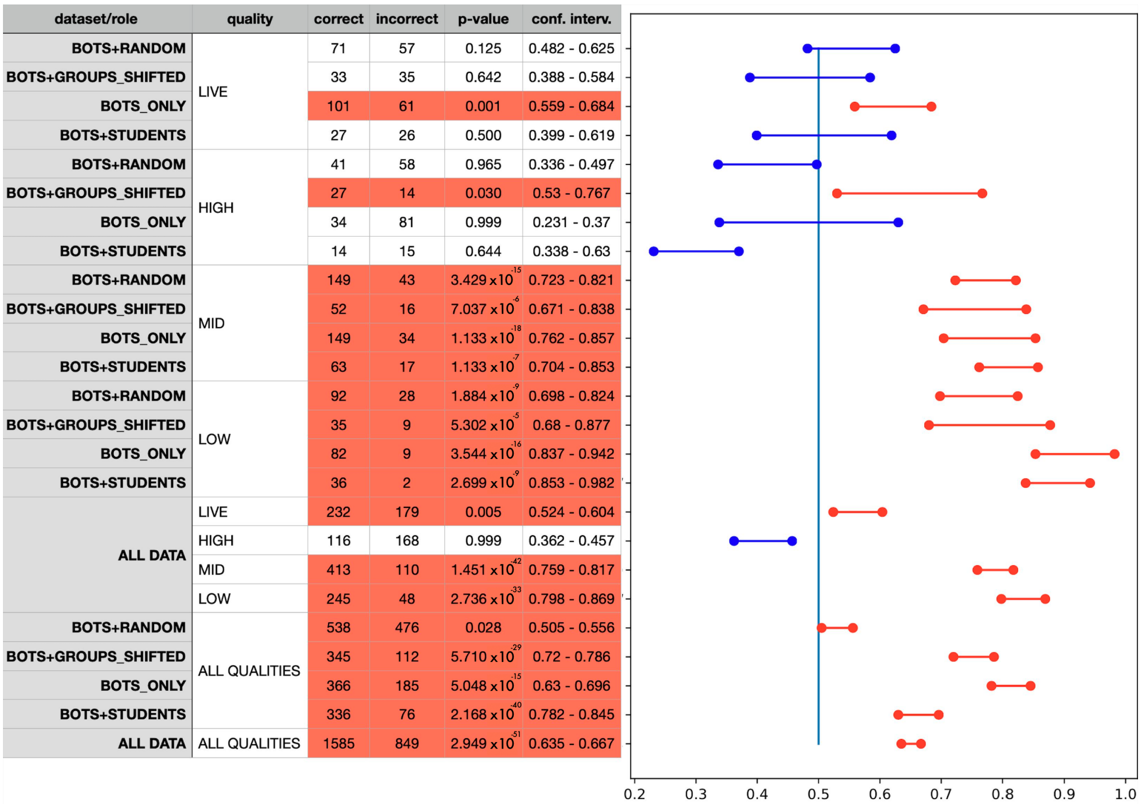
For each annotator answer, we save the record that includes identifiers of the annotator and analysed account, real label (bot/user), and annotator answer (bot/user/IDN).

We calculate the number of successes for hypotheses tests and efficiency metrics for the hypothetical classifier based on the difference between the real label and annotator answer.

4. Results and Discussion

Sixty-two students of Bonch-Bruevich University who had lectures dedicated to social media security and bot detection participated as annotators in the experiment. Overall, we obtained 2877 answers. We filtered "I do not know" answers (left only those answers in which the annotators were sure) and obtained 2434 answers. In order to detect unscrupulous annotators, we have also analysed time spent on labelling by annotators and their ratio of "I do not know" answers. As a result, all labels of two unscrupulous annotators were removed due to too small response time in comparison with other annotators.

The results of the experiment are presented in Figure 3. The results grouped by the user roles and bot quality are presented in the first and second columns. The table also shows the number of correct and incorrect answers for each sub-sample of answers, *p*-value, and confidence intervals. We also highlight rows with *p*-value < 0.05 with red colour (scenarios when annotators can recognise bots). In the confidence interval chart, we draw a vertical blue line corresponding to a random classifier with the probability of success = 0.5. Google colab notebook with statistical analysis is also presented in our dataset MKVKTT21 [27].



dataset/role	quality	correct	incorrect	p-value	conf. interv.
BOTS+RANDOM	LIVE	71	57	0.125	0.482 - 0.625
BOTS+GROUPS_SHIFTED		33	35	0.642	0.388 - 0.584
BOTS_ONLY		101	61	0.001	0.559 - 0.684
BOTS+STUDENTS		27	26	0.500	0.399 - 0.619
BOTS+RANDOM	HIGH	41	58	0.965	0.336 - 0.497
BOTS+GROUPS_SHIFTED		27	14	0.030	0.53 - 0.767
BOTS_ONLY		34	81	0.999	0.231 - 0.37
BOTS+STUDENTS		14	15	0.644	0.338 - 0.63
BOTS+RANDOM	MID	149	43	$3.429\,x10^{-15}$	0.723 - 0.821
BOTS+GROUPS_SHIFTED		52	16	$7.037\,x10^{-6}$	0.671 - 0.838
BOTS_ONLY		149	34	$1.133\,x10^{-18}$	0.762 - 0.857
BOTS+STUDENTS		63	17	$1.133\,x10^{-7}$	0.704 - 0.853
BOTS+RANDOM	LOW	92	28	$1.884\,x10^{-9}$	0.698 - 0.824
BOTS+GROUPS_SHIFTED		35	9	$5.302\,x10^{-5}$	0.68 - 0.877
BOTS_ONLY		82	9	$3.544\,x10^{-16}$	0.837 - 0.942
BOTS+STUDENTS		36	2	$2.699\,x10^{-9}$	0.853 - 0.982
ALL DATA	LIVE	232	179	0.005	0.524 - 0.604
	HIGH	116	168	0.999	0.362 - 0.457
	MID	413	110	$1.451\,x10^{-42}$	0.759 - 0.817
	LOW	245	48	$2.736\,x10^{-33}$	0.798 - 0.869
BOTS+RANDOM	ALL QUALITIES	538	476	0.028	0.505 - 0.556
BOTS+GROUPS_SHIFTED		345	112	$5.710\,x10^{-29}$	0.72 - 0.786
BOTS_ONLY		366	185	$5.048\,x10^{-15}$	0.63 - 0.696
BOTS+STUDENTS		336	76	$2.168\,x10^{-40}$	0.782 - 0.845
ALL DATA	ALL QUALITIES	1585	849	$2.949\,x10^{-51}$	0.635 - 0.667

Figure 3. Hypotheses testing results—for sophisticated bots, user results are similar to random guesses.

From the results, we can draw the following conclusions:

1. According to the last row, for Hypothesis 1, we reject H_0 as we see that humans can detect bots with the probability of success ≈0.65;
2. According to the quality blocks, for Hypothesis 2, we reject H_0 for simple bots with MID and LOW quality because humans' bot detection ability is above 0.5. In almost all tests for HIGH and LIVE qualities, we can not reject H0, so we conclude that humans can not detect sophisticated bots. There are two exceptions, namely, BOTS_ONLY scenario with LIVE quality and BOTS+GROUPS_SHIFTED scenario with HIGH quality, but their lowest border is close to random =0.5;

3. According to the dataset ALL DATA, we reject H_0 only for HIGH quality, while LIVE quality is very close to 0.5;
4. According to the dataset ALL QUALITIES, we cannot reject H_0, but notice that for the BOTS+RANDOM scenario, detection ability is very close to 0.5.

As we can see from experiment results in Figure 3, the human ability of bot detection varies and depends on the bot detection scenario and bot type. Humans can generally detect LOW and MID quality bots, while HIGH and LIVE quality bots have become too sophisticated, and users' answers are nearly random.

Table 3 depicts theoretical classifier efficiency measures.

Table 3. Hypothetical classifier testing results—note that for real classifier efficiency would be even lower.

	ALL DATASETS			BOTS+RANDOM			BOTS+GROUPS_SHIFTED			BOTS+STUDENTS		
	BACC	**TPR**	**TNR**	**BACC**	**TPR**	**TNR**	**BACC**	**TPR**	**TNR**	**BACC**	**TPR**	**TNR**
ALL QUALITIES	0.647	0.666	0.627	0.522	0.655	0.389	0.752	0.665	0.839	0.812	0.700	0.925
LOW	0.732	0.836		0.578	0.767		0.817	0.795		0.936	0.947	
MID	0.708	0.790		0.582	0.776		0.802	0.765		0.856	0.788	
HIGH	0.518	0.408		0.401	0.414		0.749	0.659		0.704	0.483	
LIVE	0.595	0.564		0.472	0.555		0.662	0.485		0.717	0.509	

If we analyse the results of hypothetical classifiers evaluation (Table 3), we can see that annotation of training datasets by humans leads to the following consequences:

1. Probably, such bot detection systems can not detect sophisticated bots. HIGH and LIVE quality TPR varies from ≈0.5 to ≈0.6. It means that the TPR of a real system may be approximately two times lower than expected and may even equal to FPR in some scenarios, which makes such a system useless (as TPR is multiplied with errors of trained machine learning model, which in most cases according to [16] is ≈0.9);
2. For LOW and MID quality, the real system's TPR may be enough to detect a sufficient number of bots to spot the attack; however, the accuracy of such a system would be lower than stated;
3. On the one hand, training on a representative sample allows the model to remember patterns of the genuine-users' normal behaviour. On the other hand, we see how dramatically TPR and TNR for the representative sample (BOTS+RANDOM) decrease. It means that in the BOTS+RANDOM scenario, the real system's "found bots output" would contain more genuine-users than bots, which makes it impossible to use such systems.

Generally, we can expect that a system trained on a human-annotated dataset is applicable only in scenarios when users in an analysed sample have strong homophily. Therefore, supervised bot detection models can be used only for analysing specific communities and should be retrained when applied to another one. In other scenarios, real TPR and balanced accuracy can be up to two times less than expected, which leads to the incorrect detection of bot attacks, especially the most complicated attacks in which top-quality bots are common.

We also point out the main limitation of our study—the experiment was conducted on the VKontakte social network and Russian bot-market segment. Therefore, our conclusions should be tested separately for other social networks and bot-traders because (as noted in [31]) Russian bot service providers dominate the social media manipulation market. In other social networks, bots may be less sophisticated, and our results may be less relevant until techniques for creating sophisticated bots are not spread to other platforms.

It is important to note that the annotation process can be improved using more complex pipelines, resulting in more accurate labels. For example, a researcher may select annotators with better bot recognition skills, which would also require a ground-truth dataset. Alternatively, the labelling process may involve multiple answers for one account from different annotators, similar to an ensemble of weak learners. However, these approaches could reduce the diversity of the resulting sample, as there would be fewer annotators or fewer labelled accounts due to the need for several annotators to analyse the same accounts. The efficiency of such approaches should be studied separately, but it is expected that they will not be able to achieve the same level of accuracy as the purchase method, particularly when detecting sophisticated bots. To reduce the effect of the experimental bias in our case, we added a labelling option "I don't know" and only used labels in which users were sure.

Nonetheless, the research presented in this study highlights the potential drawbacks of utilising human-labelled annotation methods for training supervised bot detection systems, and suggests that the effectiveness of machine learning solutions trained on such datasets should be viewed with caution.

5. Conclusions

We tested a number of hypotheses that demonstrate the existence of the "ground-truth label problem" in supervised bot detection systems. The results showed that humans could not detect sophisticated bots (HIGH & LIVE quality) in most scenarios—bots were detected in 2/8 scenarios with BACC $\approx$ 0.6. At the same time, humans have some ability to detect simple bots (MID & LOW quality)—detected in 8/8 scenarios with BACC $\approx$ 0.8. Therefore, using human labelling for forming a training dataset may decrease the system's bot detection efficiency up to two times compared to the efficiency measured in model validation—drop down of BACC to $\approx$0.5 in worst cases. We also expect that such systems may detect bots only in specific scenarios—inside a community with strong homophily (BACC $\approx$ 0.7–0.9), and may not detect bots when analysing a random sample of social network accounts (BACC $\approx$ 0.4–0.5). The obvious solution is to abandon the human annotation techniques for generating datasets and use more reliable labelling methods, such as purchasing bots directly from bot-traders.

Author Contributions: Conceptualization and study design, M.K.; software development, M.K.; software validation, O.T.; statistical analysis, M.K. and O.T.; original draft preparation, M.K.; review and editing, O.T. and V.D.; data acquisition, M.K. and L.V.; supervision and funding acquisition, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets MKVKTT21 and MKVK21 were analysed in this study. This data and code examples can be found here: https://github.com/guardeec/datasets (accessed on 1 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dialektakis, G.; Dimitriadis, I.; Vakali, A. CALEB: A Conditional Adversarial Learning Framework to Enhance Bot Detection. *arXiv* **2022**, arXiv:abs/2205.15707. [CrossRef]
2. Cresci, S.; Petrocchi, M.; Spognardi, A.; Tognazzi, S. Better Safe Than Sorry: An Adversarial Approach to Improve Social Bot Detection. In Proceedings of the 10th ACM Conference on Web Science, Boston, MA, USA, 30 June–3 July 2019.
3. Shi, P.; Zhang, Z.; Choo, K.K.R. Detecting Malicious Social Bots Based on Clickstream Sequences. *IEEE Access* **2019**, *7*, 28855–28862. [CrossRef]
4. Cresci, S. A decade of social bot detection. *Commun. ACM* **2020**, *63*, 72–83. [CrossRef]
5. Kolomeets, M.; Chechulin, A. Analysis of the malicious bots market. In Proceedings of the 2021 29th Conference of Open Innovations Association (FRUCT), Tampere, Finland, 12–14 May 2021; pp. 199–205.
6. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. *Commun. ACM* **2016**, *59*, 96–104. [CrossRef]
7. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 963–972.
8. Beskow, D.M.; Carley, K.M. Bot-Match: Social Bot Detection with Recursive Nearest Neighbors Search. *arXiv* **2020**, arXiv:abs/2007.07636.

9. Tan, Z.; Feng, S.; Sclar, M.; Wan, H.; Luo, M.; Choi, Y.; Tsvetkov, Y. BotPercent: Estimating Twitter Bot Populations from Groups to Crowds. *arXiv* **2023**, arXiv:abs/2302.00381.
10. Yang, C.; Harkreader, R.; Gu, G. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Trans. Inf. Forensics Secur.* **2011**, *8*, 1280–1293. [CrossRef]
11. Alsubaei, F.S. Detection of Inappropriate Tweets Linked to Fake Accounts on Twitter. *Appl. Sci.* **2023**, *13*, 3013. [CrossRef]
12. Dubasova, E.; Berdashkevich, A.; Kopanitsa, G.; Kashlikov, P.P.; Metsker, O. Social Network Users Profiling Using Machine Learning for Information Security Tasks. In Proceedings of the 2022 32nd Conference of Open Innovations Association (FRUCT), Tampere, Finland, 9–11 November 2022; pp. 87–92. [CrossRef]
13. Tyagi, T.; Sharma, P.; Bansal, R.; Anjali; Jain, K.; Bansal, P.; Malik, K. Twitter Bot Detection using Machine Learning Models. In Proceedings of the 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 19–20 January 2023; pp. 26–30.
14. Hays, C.; Schutzman, Z.; Raghavan, M.; Walk, E.; Zimmer, P. Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection. *arXiv* **2023**, arXiv:2301.07015.
15. Cresci, S.; Di Pietro, R.; Spognardi, A.; Tesconi, M.; Petrocchi, M. Demystifying Misconceptions in Social Bots Research. *arXiv* **2023**, arXiv:2303.17251.
16. Orabi, M.; Mouheb, D.; Al Aghbari, Z.; Kamel, I. Detection of bots in social media: A systematic review. *Inf. Process. Manag.* **2020**, *57*, 102250. [CrossRef]
17. Igawa, R.A.; Barbon, S., Jr.; Paulo, K.C.S.; Kido, G.S.; Guido, R.C.; Júnior, M.L.P.; da Silva, I.N. Account classification in online social networks with LBCA and wavelets. *Inf. Sci.* **2016**, *332*, 72–83. [CrossRef]
18. Jr, S.B.; Campos, G.F.; Tavares, G.M.; Igawa, R.A.; Jr, M.L.P.; Guido, R.C. Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–17. [CrossRef]
19. Dickerson, J.P.; Kagan, V.; Subrahmanian, V. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 17–20 August 2014; pp. 620–627.
20. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. Fame for sale: Efficient detection of fake Twitter followers. *Decis. Support Syst.* **2015**, *80*, 56–71. [CrossRef]
21. Kolomeets, M.; Chechulin, A.; Kotenko, I.V. Bot detection by friends graph in social networks. *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.* **2021**, *12*, 141–159.
22. Subrahmanian, V.S.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F. The DARPA Twitter bot challenge. *Computer* **2016**, *49*, 38–46. [CrossRef]
23. Morstatter, F.; Wu, L.; Nazer, T.H.; Carley, K.M.; Liu, H. A new approach to bot detection: Striking the balance between precision and recall. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 533–540.
24. Echeverria, J.; Zhou, S. Discovery, retrieval, and analysis of the 'star wars' botnet in Twitter. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, 31 July–3 August 2017; pp. 1–8.
25. Kantepe, M.; Ganiz, M.C. Preprocessing framework for Twitter bot detection. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 630–634.
26. Gallwitz, F.; Kreil, M. Investigating the Validity of Botometer-Based Social Bot Studies. In Proceedings of the Disinformation in Open Online Media: 4th Multidisciplinary International Symposium, MISDOOM 2022, Boise, ID, USA, 11–12 October 2022; pp. 63–78.
27. Kolomeets, M. MKVK2021 and MKVKTT2021 Security Datasets. Available online: https://github.com/guardeec/datasets (accessed on 26 November 2023).
28. Top Websites Ranking by Country (on 1 March 2023). Available online: https://www.similarweb.com/top-websites/computers-electronics-and-technology/social-networks-and-online-communities (accessed on 26 November 2023).
29. VK Report for q1 2022 (in Russian). Available online: https://vk.com/main.php?subdir=press&subsubdir=q1-2022-results (accessed on 26 November 2023).
30. Branitskiy, A.; Levshun, D.; Krasilnikova, N.; Doynikova, E.; Kotenko, I.V.; Tishkov, A.; Vanchakova, N.; Chechulin, A. Determination of Young Generation's Sensitivity to the Destructive Stimuli based on the Information in Social Networks. *J. Internet Serv. Inf. Secur.* **2019**, *9*, 1–20.
31. *The Black Market for Social Media Manipulation*; Research Report; NATO Strategic Communications Centre of Excellence: Riga, Latvia, 2018. Available online: https://stratcomcoe.org/cuploads/pfiles/web_nato_report_-__the_black_market_of_malicious_use_of_social_media-1.pdf (accessed on 1 February 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.