The Devil's Triangle: Ethical Considerations on Developing Bot Detection Methods

Andree Thieltges, Florian Schmidt, Simon Hegelich

FoKoS, University of Siegen, Weidenauer Straße 167, 57076 Siegen, Germany
andree.thieltges@uni-siegen.de, florian.schmidt@uni-siegen.de, simon.hegelich@uni-siegen.de

Abstract

Social media is increasingly populated with bots. To protect the authenticity of the user, experience machine learning algorithms are used to detect these bots. Ethical dimensions of these methods have not been thoroughly considered yet. Taking histogram analysis of Twitter users' profile images as example, the paper demonstrates the trade-offs of accuracy, transparency, and robustness. Because there is no general optimum in ethical considerations, these dimensions form a "devil's triangle".

Bots as an ethical problem - manipulation vs. authenticity

Today one cannot imagine social interaction without internet-based social media networks and platforms: So many people have a Facebook or Twitter account and use it to interact with each other or to gather or spread information at any time (Hegelich and Shahrezaye 2015). Furthermore, one has the opportunity to create social content and share their pictures, videos and events. These features, in addition to the rapid growth of the digital communities, render them interesting not only for private use, but commercial, political and religious purposes as well. In fact, people are interested in such "social content" for many reasons and mostly consider the information that one can gather at social media platforms as original or authentic (Ratkiewcz et al. 2011). Recently, the façade of authenticity and reliable source of information began to crumble, as social bots and botnets are being exposed thereby strengthening the evidence of deception, abuse and manipulation within social networks (Ferrara et al.2015).

Exploring the techniques of how bots adulterate authenticity in social platforms, one may start with the inflation of "fake followers" which simply boost the popularity of particular Twitter accounts. This quite popular practice is associated with making

> "an account more trustworthy and influential, in order to stand out from the crowd and to attract other genuine followers" (Cresci et al. 2015; Edwards et al. 2014).

Besides this questionable increase in one's influence and enlargement of one's audience, recently the evidence for continuous "social fraud" has accumulated on different stages of social networking. As an exemplary case one might consider the marginalization of the #YaMeCance Twitter hashtag, which was a major junction point in spreading information and organizing modern Mexican protest movement against violence and corruption. Social bots undercut this hashtag by tweeting spam or anti-protest messages, drowning out real conversations with noise and following or threatening the activists (Finley 2015).

In addition to the consequence that these social activists get muzzled by employing bots on their social media hub, this sample of bot usages led also to a basic ethical dilemma: How can the gathering of original information and authentic interaction within social networks be assured? Since bots and botnets try to copy human behavior to disguise their machined actions, one can imagine this is a rather tricky but essential task for keeping the integrity with regard to all of the content of social media. Currently, there are some technical solutions, more than a few provided by the operators of the platforms themselves, to detect and control the misleading bot posts and tweets and separate them from "authentic-human" information. One more or less effective way to discriminate bots and humans within social network is to employ supervised machine learning methods. Based on the behavior of humans and bots in social networks and the learned distinction between them, the aim of these techniques is to develop "detection algorithms" or "classifiers" that can separate human users from the bots. All of these approaches give more or less accurate predictions of the probability that an unknown user of a social platform either is a bot or not, and this includes the possibility of errors. One major aspect concerning sources of errors is that human behavior in

social interaction always refers to a specific context and therefore is rather heterogeneous. So, despite the fact that the bots are continuously changing and evolving to keep their camouflage, imagine the scenario of a human cyber activist who expresses her view on a sensitive subject or simply exchanges opinion and wishes to keep herself in disguise for good reason[1]. Hence her social media account may show some features which are also associated with bot accounts, for instance unusual times for system logon, a conspicuous friend-follower-ratio, an uncommon IP-address or simply a non-human avatar. It's quite conceivable that some of the classifiers will mark this human as a bot. So, the ethical dilemma mentioned above continues somehow: One cannot be sure if the prediction made by a machine learning algorithm is right all the time. Thus, not only the result of a machine learning technique to detect bots is ethically questionable, but also the decisions that will lead to it. In the following we explore the ethical content of a bot detection development process that employs a machine learning method by following up on questions about its accuracy, its transparency and its robustness.

Histogram analysis as a bot detection method

The basic idea arising from our continuous "bot hunting" projects is this: Exploring new detection methods to fit the dynamic mutability of social bots on Twitter, we diagnosed that bots often use comic characters or icons (e.g. the egg icon that Twitter provides) as their profile picture or avatar. This finding raised the question of whether it is possible to set up a framework that will scan Twitter-users by their avatar and give a prediction whether they are bots or human. Based on our verified bot and botnet database of Twitter-accounts that Simon Hegelich (2016) found during his study of Ukrainian/Russian social bots, we downloaded all of the 1948 profile pictures of the bot accounts and extracted and scaled them to the same size (400px x 400px). This ensures an increased comparability, since not all of the pictures and icons are square-cut. To complete our data set we continued this procedure with 2700 Twitter-account profile pictures from a different database where all of the users are verified humans (like some of the most active celebrities on Twitter). Subsequently, we programmed an algorithm which converted every profile picture from RGB color model into greyscale and afterwards extracted the histogram of the picture using Open CV. In a second step we divided our histogram data set into two randomized samples:

[1] There are now many known cases of activists who are prosecuted for their comments or speeches on social media platforms by state authorities. A current example may be the arrests of social media activists in the Hong Kong uprising (Zhang 2015).

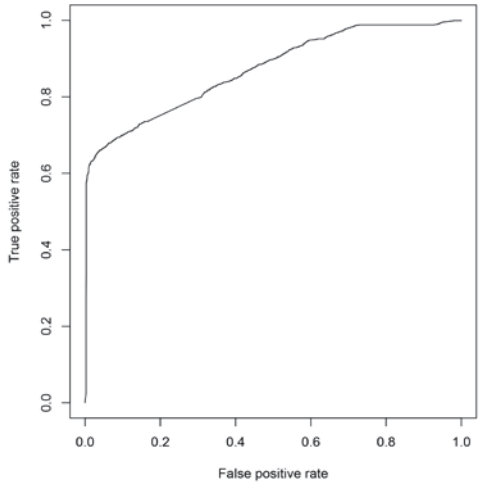

Figure 1: ROC Curve of the support vector machine classifier based on the histogram analysis of Twitter profile pictures

The smaller sample (1000 items) became our training data set (train-data), the larger one (3648 items) became our test data set (test-data). We then continued with a supervised learning method in which we taught our machine that there are bot and human accounts according to the histograms associated with each of the items in the train-data set. Based on this separation we let it compute a support vector machine classifier for the divisions of bots and humans. Finally, we let our support vector machine classifier run on our test-data set and as a result achieved 83.6% accuracy in the separation of bot and human profile picture histograms.

Ethical questions in developing bot detection by machine learning algorithms aka. the devil's triangle

Although the probability of 83.6% for a true prediction about the profile picture belonging to a bot (TRUE positive) or a human (TRUE negative) does not seem so bad at first glance, there are still sources of error which cause the algorithm to detect either bots as humans (FALSE positive) or humans as bots (FALSE negative). As aforementioned, one of the basic problems of feature-based bot classifiers is that their detection results are strongly reliant on the contextual framework of the social interaction. For instance, if you employ our "avatar classifier" on users that tweet within the context of the #indicomics hashtag instead of the #euromaidan or #maidan hashtags you are likely to get different detection results for both humans and bots. Since machine learning classifiers are based on observations of behavior and

encode this behavior into features, it is possible to expand the feature space and combine them into a more complex classifier which will increase the quantity of the bot detection and lower the number of the wrongly detected humans. One may do this by taking into account the friend-follower-ratio, the number of tweets and/or re-tweets, the duration and comparison of the access time, or even the users' meta data. From the ethical point of view, the decision to create a more complex classifier is the first step into the "devil's triangle":

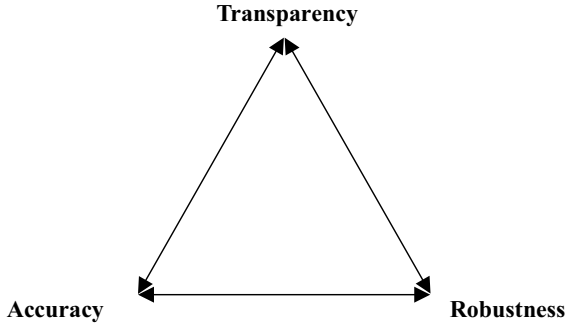

Figure 2: The devil's triangle

Coming back to the social net activist who disguises herself, a classifier that combines certain features will not detect her as a human user, but carry on to treat her as a bot, with the result that her access to the social platform will be restricted or she will be marked as a bot. She cannot explain herself why this happens, because she cannot reconstruct the features that the algorithm uses to differentiate machines and humans.

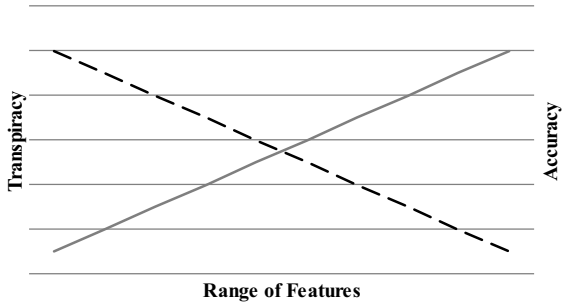

Figure 3: Increase of the features range will enhance the accuracy and decrease the transparency.

In other words

> "it will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also transparent to inspection – to name one of many socially important properties" (Bostrom and Yudkowski 2014).

Dealing with transparency

But what if one opens the black box of machine learning algorithms which are employed to detect bots and botnets? Imagine that somebody who has developed a classifier to detect bots (whether it is a complex neural network or a simple decision tree-based approach) is going to publicize her method, features and results.

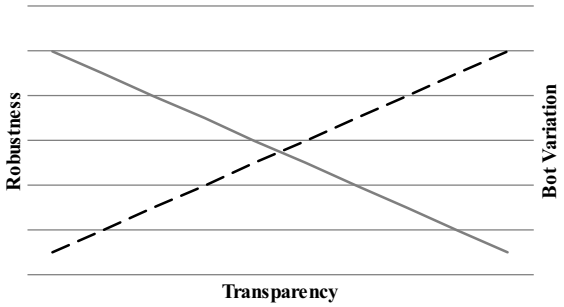

Figure 4: Increased transparency will cause continuous bot variations and a decrease in detection robustness.

Not long before, bot-operators will adapt their machines so this detection algorithm becomes invalid. Speaking in ethical terms, (too much) transparency in this case will increase the uncertainty and the probability of manipulation and fraud within social networks and put it on a whole new level, because a generation of new bots (referred to as 2.0) comes into being.

Regarding our triangle, the classifier only detects the "bot-1.0" generation but cannot detect the "bot-2.0" generation, hence the number of FALSE positive and FALSE negative predictions will increase. Furthermore, if a new classifier is developed to identify the new bot-2.0 generation, it must include all the relevant features to separate three classes of users (humans, bots-1.0 and bots-2.0) instead of two. This represents an intrinsic source of error, since there is a permanent danger to over- or underfit the prediction model which is the base for the new classifier. Eventually, the larger variance in the bot population will cause another threat to our prediction: The fading of the robustness against manipulation.

Variance vs. robustness – back to the beginning

To detect the newly formed bot generation will require an expanded classifier, since the one that was developed in our example only identifies the 1.0-bots. Moreover, the detection of these bots will also diminish, in fact they will also alter their behavior based on their advanced knowledge of the bot detection algorithm. There will be a larger variation in features that will work to identify bots in the future, since not all of the generation 2.0-bots will behave similarly. To set up an algorithm which proves to be robust, one could probably proceed to specify the advanced features of the new bot generation. This may result in numerous "deductive" approaches. Or one could try to continue with a more generic approach, by finding universal features of the bot population. Both ways may be

problematic according to the ratio of error one will achieve: If one will develop a specialized algorithm it presumably has a low rate of errors by identifying bots which apply to the particular set of features. If one employs this algorithm to a broader data-set i.e., the variation of the bot features may increase and so will the sources of error. If one computes an algorithm based on more universal features, it may achieve a rather good rate of error by detecting bots in general. But what if one reduces the set of data to a set where only the bots remained that specified their behavior? One can comprehend that both approaches cause a higher accuracy if they employ either a broader or a smaller set of data. Finally, we return to our zero point of the "devil's triangle", starting once again by seeking an algorithm that will increase the accuracy of our bot detecting method and, by doing so, reduce the manipulation in social networks.

Conclusion

One might get the notion that bringing ethical aspects into machine learning techniques for bot detection leads to a never ending task: Every solution to strengthen our rate of error, the transparency or the robustness of a given algorithm somehow opens up another stage of ethical problem, in other words, whatever one does is wrong. One will get into trade-offs, since the reciprocal problems with accuracy, transparency and robustness cannot be solved and moreover, one cannot weight them against each other. We showed that alternative behavior of both, human and bot users of social platforms will get the detection methods into trouble. Following the idea of alternative behavior, one may approach the developing process of machine learning techniques for bot detection under aspects of bounded rationality: The observation of human and bot behavior within social platforms could be considered as a "set of behavior alternatives" that will provide reference alternatives in choices and decisions. To specify these behavioral alternatives, one may employ a

> "subset of behavior alternatives" that the organism "consider" or "perceive". That is, the organism may make its choice of alternatives within a set of more limited than the whole range objectively available to it."(Simon 1955)

This may cause the implementation of features that will respect the variety of human behavior instead of isolating the statistical mainstream. However, developers of machine learning bot detection methods should keep in mind, that the – yet currently still simple – ethical problems will remain, and that there is no "optimization under constraints" (Gigerenzer and Selten 2002).

Nevertheless, bot detection methods have to deal with these constraints and this may be the major task in developing a bot detection algorithm. It may sound simple but in our opinion, the key suggestion for such a developing process is to clarify the detection goals. To come to a decision, our "devil's triangle" may be a - still quite simple - tool regarding ethical considerations and dealing with the aforementioned trade-offs. For instance, if one has the goal to maintain the safety of a social network and therefore develop and employ bot detection methods, several features to discriminate the humans and the bots may be involved and the use of combined sets of features may increase the accuracy of the algorithms. But not everything that is technically feasible has to be used. In a development process one should ask if the expansion of the feature range is a good choice in regard to the purpose one wishes to achieve. Moreover, to verify or falsify the benefit of an increased feature range, every single feature of the classifier must be reviewed relating to its purpose. In addition, one has to take into account that the goal of network safety may affect other crucial issues. As we stated before, some of the ethical problems cannot be solved. To enable oneself to include ethical considerations in one's decision, one should at least be aware of those problems and this is where our "devil's triangle" may give support.

References

Bostrom, N. and Yudkowsky, E. 2014. The ethics of artificial intelligence. In The Cambridge Handbook of Artificial Intelligence, 316-334. Cambridge: Cambridge University Press

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. 2012. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? In Transactions on Dependable and Secure Computing 2012-11, 811-824. Washington D.C.: IEEE Computer Society.

Cresci, S., Di Pietro, R., Petrocchi, M.; Spognardi, A. and Tesconi, M. 2015. Fame for sale: efficient detection of fake Twitter follower. In Decision Support Systems, 80, 56-71. Elsevier B.V.

Edwards, C., Edwards, A., Spence, P.R., and Shelton, A. 2014. Is that a bot running the social media feed? Testing the differences of communication quality for a human agent and a bot agent on Twitter. In Computers in Human Behavior 33, 372-376. Elsevier B.V.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. 2015. The Rise of Social Bots. Accessed 2015-10-07. http://arxiv.org/pdf/1407.5225v3.pdf.

Finley, K. 2015. Pro-Government Twitter Bots Try to Hush Mexican Activist. In Wired Magazine. Accessed 2015-10-07. http://www.wired.com/2015/08/pro-government-twitter-bots-try-hush-mexican-activists/.

Gigerenzer, G. and Selten, R. 2002 Rethinking Rationality. In Bounded Rationality. The Adaptive Toolbox, 1-13. Cambridge, Mass.: The MIT Press.

Hegelich, S. 2016. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian/Russian Social Botnet. In

Media Bias. Tagungsband 112 der Schriftenreihe des Instituts für Rundfunkrecht. München: Verlag C.H. Beck.

Hegelich, S. and Shahrezaye, M. 2015. The communication behavior of German MPs on Twitter: Preaching to the converted and attacking opponents. In European Policy Analysis Vol 1. Number 2 – Fall 2015, 155-174. Accessed 2015-10-08. http://joom.ag/aOgp/p154#.VhYwpdkNUJg.mailto.

Ratkiewcz, J., Conover, M., Meiss, M.; Gonçalves, B., Patil, S., Flammini, A. and Menczer, F. 2011. Truthy: mapping the spread of astroturf in microblog streams. In Proceedings of the 20th international conference companion on World wide web, 249-252. New York: ACM.

Simon, H. 1955. A Behavioral Model of Rational Choice. In The Quarterly Journal of Economics, Vol. 69, No. 1., 99-118. Cambridge, Mass.: The MIT Press.

Zhang, J. 2015. Hong Kong's Activist Social Media Culture Under Threat. Activists claim that authorities are using an ambiguous ordinance in a selective crackdown on free speech. Accessed 2015-03-12. http://thediplomat.com/2015/06/hong-kongs-activist-social-media-culture-under-threat/.