SAGE Open
April-June 2019: 1–16

DOI: 10.1177/2158244019827715
journals.sagepub.com/home/sgo
SAGE

For Whom the Bot Tolls: A Neural Networks Approach to Measuring Political Orientation of Twitter Bots in Russia

Denis Stukal[1], Sergey Sanovich[1], Joshua A. Tucker[1,2,3], and Richard Bonneau[1,3,4]

Abstract

Computational propaganda and the use of automated accounts in social media have recently become the focus of public attention, with alleged Russian government activities abroad provoking particularly widespread interest. However, even in the Russian domestic context, where anecdotal evidence of state activity online goes back almost a decade, no public systematic attempt has been made to dissect the population of Russian social media bots by their political orientation. We address this gap by developing a deep neural network classifier that separates pro-regime, anti-regime, and neutral Russian Twitter bots. Our method relies on supervised machine learning and a new large set of labeled accounts, rather than externally obtained account affiliations or orientation of elites. We also illustrate the use of our method by applying it to bots operating in Russian political Twitter from 2015 to 2017 and show that both pro- and anti-Kremlin bots had a substantial presence on Twitter.

Keywords

neural network, natural language processing, social media, Twitter bots, propaganda, Russia

Introduction

Computational propaganda and the use of automated accounts in social media have been attracting an increasing amount of public attention. Both the mass media and the general public have been alarmed by evidence of bots and trolls being used at an unprecedented scale for political purposes throughout the world. Of particular interest have been reported attempts by the Russian government—consistently denied by the Kremlin (TASS, 2017)—to leverage computational propaganda tools to interfere with the electoral process in the United States and Europe (Alandete, 2017; Arnsdorf, 2017; Booth, Weaver, Hern, & Walker, 2017; Entous, Nakashima, & Jaffe, 2017; Grassegger & Krogerus, 2017; Nimmo, 2017; O'Sullivan & Byers, 2017; Popken, 2017; Rosenberg, 2018; Shane, 2017; Wells & Seetharaman, 2017).

Identifying bots and trolls on social media platforms such as Twitter is a burgeoning area of research (Bessi & Ferrara, 2016; Brachten, Stieglitz, Hofeditz, Kloppenborg, & Reimann, 2017; Chu, Gianvecchio, Wang, & Jajodia, 2012; Forelle, Howard, Monroy-Hernndez, & Savage, 2015; Hegelich & Janetzko, 2016; Miller, 2017; Ratkiewicz et al., 2011; Schfer, Evert, & Heinrich, 2017) that faces significant empirical (Cresci, Di Pietro, Petrocchi, Spognardi, & Tesconi, 2017; Gilani, Farahbakhsh, Tyson, Wang, & Crowcroft, 2017; Oentaryo, Murdopo, Prasetyo, & Lim, 2016; Stieglitz et al., 2017; Varol, Ferrara, Davis, Menczer, & Flammini, 2017) and conceptual (Gorwa & Guilbeault, 2018; Grimme, Preuss, Adam, & Trautmann, 2017; Stieglitz, Brachten, Ross, & Jung, 2017) challenges. The detection of (semi)automatically generated content, such as spam ads or "user" reviews of products and services, has long been an important field of study as well as a popular application for testing machine learning and high-dimensional statistical methods. The attribution of these tools to particular entities and political causes is exponentially more challenging

[1]New York University, Social Media and Political Participation (SMaPP) Laboratory, New York, NY, USA
[2]New York University, Department of Politics, New York, NY, USA
[3]New York University, Center for Data Science, New York, NY, USA
[4]New York University, Department of Biology, New York, NY, USA

Corresponding Author:
Denis Stukal, Department of Politics, New York University, 2nd Floor, 19 W. 4th Street, New York, NY 10012, USA.
Email: denis.stukal@gmail.com

(Brenner, 2007; Gorwa, 2017), and yet it is an essential step in studying bots' role in electoral campaigns and political communications in general. A few recent successful attempts to study the content of bot and troll activity capitalize on relatively short lists of accounts that were leaked by their creators or revealed through investigations and thus could plausibly be taken as "ground truth" for attribution purposes (Keller, Schoch, Stier, & Yang, 2017; King, Pan, & Roberts, 2017; Sobolev, 2018; Stewart, Arif, & Starbird, 2018; Zannettou et al., 2018).

In this article, we propose a method relying on neural networks to establish the political affiliation of bots at scale and without any prior knowledge of social media accounts' connection to a particular institution or ideology. As part of the verification of this method, we also demonstrate the feasibility of identifying a clear communication strategy amid seemingly diverse bot activities once bots are grouped by political affiliation. This opens exciting possibilities both for comparing bots with other tools of political communication and for recovering political strategies from bot activity.

Our previous research focused on identifying political bots in the Russian political Twittersphere and demonstrated that in 2014-2015, bots accounted for a surprisingly large proportion of Twitter activity (Stukal, Sanovich, Bonneau, & Tucker, 2017). More specifically, we found that on most days more than half of the accounts tweeting in our collection of Russian-language tweets about politics were bots.[1] However, in related but preliminary (qualitative and small-*n*) analysis of the uncovered bots, we discovered that Russian Twitter bots were not necessarily all pro-regime (Sanovich, Stukal, & Tucker, 2018). Instead, we repeatedly found evidence of neutral bots tweeting news headlines, as well as anti-Kremlin bots spreading information critical of Vladimir Putin. Moreover, the opposition to the Kremlin came from two distinctly different vantage points: opposition to the state of Russian domestic politics, and opposition to Russia's actions vis-à-vis the Ukrainian conflict.

Here, we propose a method to systematically analyze political orientation (or, put another way, the sustained sentiment) of Twitter bots based on the content of their tweets. We build a deep feedforward neural network (multilayer perceptron [MLP]) that uses a wide range of textual features including words, word pairs, links, mentions, and hashtags to separate four contextually relevant types of bots: pro-Kremlin, neutral/other, pro-opposition, and pro-Kiev.

Although the primary purpose of this article is methodological—to introduce and validate a replicable, supervised machine learning method for coding the political orientation of Twitter bots—we also present a number of novel empirical observations. First, we find that across our four categories of pro-Kremlin, neutral, pro-opposition, and pro-Kiev, a plurality of Russian Twitter bots in 2015-2017 were pro-Kremlin. Perhaps more surprisingly, though, is the fact that when we combine the pro-opposition and pro-Kiev bots, we find approximately as many "anti-Kremlin" bots as pro-Kremlin ones. However, the pro-Kremlin bots were more active, producing significantly more content in terms of the number of total tweets than the "anti-Kremlin" bots.

The article is organized as follows. The following section provides a detailed description of the MLP model. The next two sections describe our data and detail our orientation-classification results. We then present additional robustness checks and comparisons that allow us to go deeper into the content and attributes of ideological/orientation account clusters.

Method

Classifying Twitter bots on the basis of their political orientations is a two-step process that involves, first, separating bots from humans and, second, distinguishing between bots with different political leanings. A complication here is that it is not necessarily the case that pro-regime bots are systematically different in their activity patterns from anti-regime ones. Either group of bots could be involved in tweeting repeatedly the same text, or retweeting other Twitter users, or tweeting every *k* seconds, and so on. Moreover, there are no prior theoretical reasons to expect that bots and humans with similar political orientation would necessarily use different text or retweet different online materials or users. Thus, bot detection and their sentiment analysis are two substantively and computationally separate tasks that we address with two different classifiers.

We define bots as fully automated accounts and use a taxonomy of Twitter accounts, described in detail in Sanovich et al. (2018), that helps to distinguish bots from humans and a number of other types of accounts. The classification system provides a restrictive definition of bots, as it excludes paid human trolls (they fall under the "human" category in the taxonomy), official institution accounts (e.g., maintained by media organizations and political parties), cyborgs (i.e., accounts with both automated and manually posted content), and accounts used for the purposes of commercial spamming.

Our approach to bot detection is the same as in Stukal et al. (2017) and builds on the vast literature on ensembles of classifiers (Dietterich, 2000; Zhou, 2012). The detection algorithm relies on a "majority–unanimous" voting ensemble. The approach involves training four different component classifiers (support vector machine [SVM], ridge regression, extreme gradient boosting tree, and AdaBoost) on a set of over 40 features that characterize account tweeting activity and meta-data with fivefold cross-validation. It is a unanimous voting rule in so far as all four classifiers have to predict the account to be a bot for it to be coded as a bot. However, each classifier is run on 10 different training sets, so an account is only classified as a bot if a majority of these 10 training sets produces unanimous agreement across the four classifiers that the account was a bot. Thus, the algorithm is a "majority–unanimous" ensemble that—through

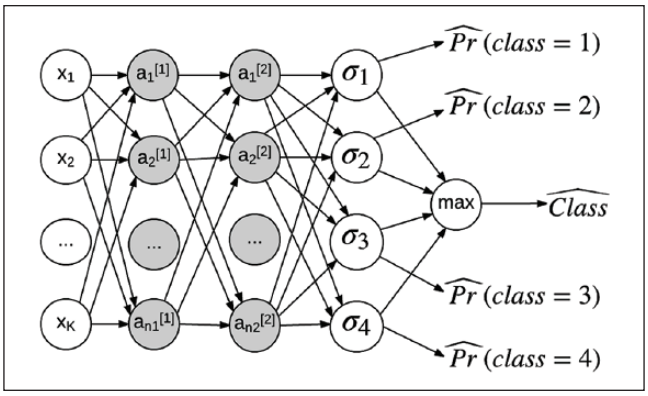

Figure 1. Multilayer perceptron used to classify bot orientation. *Note.* x_j refers to the jth input feature (here input features are tweet text, links, hashtags and mentions); $a_1^{[2]}$ refers to the first hidden unit in the second hidden layer; σ_j refers to the softmax probability of class j.

the unanimity component—produces a conservative bot-detection tool with almost perfect precision and a reasonably high recall.[2]

To detect the political orientation of bots, however, we employ a different approach. Here, we rely on the text contained in bots' tweets, including mentions (i.e., the handle of another Twitter user), hashtags, and posted links. The mapping between textual features and political orientation is a complex function generated by a diversity of possible strategies and external events. For example, a pro-regime bot could retweet a tweet from an opposition account to criticize it, but an anti-regime bot could retweet that same opposition tweet to express support. A bot could use a hashtag to either promote the message associated with it, or to swamp the hashtag with irrelevant or antagonistic content. Due to this complexity, we construct a feedforward neural network (multilayer perceptron, MLP), building on the theoretical result that sufficiently deep MLPs can approximate any complex function, including the multimodal mappings generated by the expected complex processes outlined above (Cybenko, 1989; Hornik, Stinchcombe, & White, 1989).[3]

Due to the computational complexity of training MLPs, we split the labeled set into training (80%), development (10%), and test (10%) sets instead of performing cross-validation. The general structure of the MLP is shown in Figure 1.

We follow current literature on deep neural networks (DNNs) and choose to use rectified linear unit (ReLU) activations for hidden layers (Glorot, Bordes, & Bengio, 2011) as they do not saturate, which speeds up the training process and mitigates the vanishing gradient problem (Goldberg, 2017). Hence, all hidden units here are ReLU activations applied to a linear function of the activations from the previous layer:

$$a^{[l]} = max(0, \mathbf{W^{[l]}} \times a^{[l-1]} + b^{[l]}),$$

where $a^{[l]}$ is an activation in layer l, $l \in \{\text{input}, 1, ..., L\}$, $\mathbf{W^{[l]}}$ is the weights matrix, and $b^{[l]}$ is the bias (constant) term in layer l. As one can see, a ReLU activation outputs a linear function of the activations from the previous layer, unless that function is negative, in which case the output is 0. Hence, it is a very simple nonlinear transformation that has, however, proven to be highly effective and efficient in learning DNN parameters (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015; Szegedy et al., 2015).

To avoid overfitting, we use the softmax cross-entropy cost function with L_2-regularization for the weights:

$$\mathfrak{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\prod_{k=1}^{K}\left(\frac{\exp\left(\mathbf{w_k^{[L]}} \times a^{[L-1]} + b_k^{[L]}\right)}{\sum_{m=1}^{K}\exp\left(w_m^{[L]} \times a^{[L-1]} + b_{\boldsymbol{m}}^{[L]}\right)}\right)^{1\left(y_i \in C_k\right)}\right) + \lambda\sum_{l=1}^{L} \| \mathbf{W^{[l]}} \|_F^2,$$

where y_i is the observed label for observation i; C_k denotes kth class; $a^{[L-1]}$ refers to the activations from the last hidden layer; $\mathbf{w_k^{[L]}}$ and $b_k^{[L]}$ are, respectively, the weights vector and the bias scalar for the kth label in the output layer; $\mathbf{W^{[l]}}$ is the weights matrix in layer l, and $\| W^{[l]} \|_F$ is its Euclidean (Frobenius) norm; $1(y_i \in C_k)$ is the indicator function that equals 1 if $\mathbf{i}$'s observed label is k, and 0 otherwise.

We experiment with the network architecture and hyperparameters by searching the best model specification via grid search over different numbers of hidden layers, hidden units, and L_2 penalty values. In particular, we considered two-layer neural networks with 10, 50, and 100 hidden units and three-layer networks with 10-3, 10-5, 50-10, 50-20, 100-10, 100-20, and 100-50 hidden units,[4] while L_2 penalty values ranged from 10^{-4} to 0.01 (in total, 80 model versions). Each model parameterization was evaluated based on its accuracy computed on the development set. We selected the model with the best performance on the development set and then provided its final performance evaluation using the test set (reported in Table 2). The model outputs predict probabilities for every class and assign every observation to the class with the largest predicted probability. To focus our discussion on our top model predictions, we only consider unlabeled bots with the largest predicted probability above .9. We report results for the .7 prediction probability threshold to provide an additional robustness check in Supplemental Appendix A.

Data Collection and Coding

In this article, we rely on approximately 38 million tweets collected using the Twitter Streaming API from 2015 to 2017 using a list of 86 keywords and hashtags that refer to different aspects of Russian politics as discussed on Twitter

(see the full list in Supplemental Appendix B). The list included both permanently relevant words (e.g., surnames of major politicians: Putin, Medvedev, Nemtsov, Navalny, etc.) and situationally used hashtags (such as used by Putin supporters during his televised "direct line" with citizens or used by the opposition during big rallies in Moscow). While we have periodically updated the list of keywords, it did not encompass the entire universe of politically relevant keywords. However, as it included the most generic words used in most tweets about politics (such as names of parties and politicians), we were able to collect a broadly representative sample of political speech in Russian Twitter.

Previous research shows that bots avoid using geotagged tweets, which is why, in order not to miss bot-generated tweets, we did not restrict our collection to geolocated data. This approach, however, poses the additional problem of separating the Russian Twittersphere from the communities of Twitter users in other countries. We addressed this problem by considering only those accounts that used Russian as the account interface language with at least 75% of their tweets (as recorded in those tweets' meta-data).

In addition, we restricted our attention to accounts with at least 10 political tweets each year between 2015 and 2017. With these restrictions, we end up with almost 11 million tweets from more than 20,000 accounts. As we are interested in predicting the political orientation of bots, we applied our MLP model to accounts predicted to be bots by the bot-detection algorithm described in Stukal et al. (2017) and, for reasons that we lay out in the next section, focused on accounts that tweeted at least 100 times. This left us with a data set of 5,047 bot accounts with at least 100 tweets' each.

As explained in the previous section, our orientation classifier is a supervised learning method that requires a labeled set. When creating the labeled set, we used static Twitter snapshots that represent what an account would have looked like if it only contained data from our collection. Thus, the coding process was both consistent with the actual data we analyzed afterward and reproducible.[5] The Twitter snapshots contain both the text of the tweets and the account meta-data (user bio, number of followers, and other information ordinarily displayed on Twitter) that are available from collected tweets. If a tweet was deleted after being scraped into our collection, it still remained in the collection, but the Twitter snapshot showed it as plain text, as opposed to the typical Twitter style. In this way, our coders were able to view something that resembles a Twitter account online, but in a manner that only contains the tweets in our collections and therefore is completely reproducible (as opposed to, for example, asking the coders to load the Twitter account live online, which would result in a different display depending on the time of the coding).

To ensure high coding and reproducibility standards for the labeled set, we recruited seven Russian native speakers, all undergraduates in social sciences, as coders. They received detailed written instructions on how to classify Twitter accounts by type and code their political orientation, and then undertook a trial round of coding. We reviewed the results and had a Skype training session with the coders focusing on the most common, as well as individual, mistakes. This was followed by another training round of coding with feedback provided in written form. By that time, all the coders demonstrated sufficient proficiency in classification, and we proceeded with the actual coding.

Supplemental Appendix C presents the instructions (translated from Russian) used by coders to both classify Twitter accounts by type and code their political orientation. Supplemental Appendices D and E present the classification algorithms for both tasks in a schematic form. Supplemental Appendix F presents screenshots of typical accounts in each category and additional tips for distinguishing between them. While the typology of Twitter accounts (bots, humans, cyborgs, etc.) we created was meant to be as generalized as possible, the specific examples, typical mistakes, and distinguishing characteristics we provided to coders were empirical and contextual, that is, developed when we were manually going through the collection trying to balance two competing goals: (a) every account the coders might encounter having a category whose description it matches closely, and (b) limiting the number of categories to facilitate the analysis. As a result, our typology includes 13 terminal categories, of which seven are various types of bots. As mentioned above, our definition of bots is conservative and includes only accounts consisting entirely of content automatically lifted from external sources. However, our instructions and algorithms were designed to minimize complex judgment about the nature of the account and, instead, guide coders toward assigning a specific category based on its easily identifiable characteristics. For example, an anonymous account featuring same-size images, all related to news events, posted every few minutes, and lacking any other type of tweets would naturally go into "bot with pictures" category. Although our instructions contain guidance for making a decision in less obvious cases, it is important to keep in mind that they constitute a small fraction of our data set and hence our training set. Further details of our coding process for account type are available in Stukal et al. (2017).

After determining the type of Twitter account, our coders were asked to identify the account's political orientation. Building on our preliminary analysis in Sanovich et al. (2018), we asked the coders to label each account as pro-Kremlin, pro-opposition, pro-Kiev, or neutral; this last category was defined as a residual category for any account that was not coded as belonging to one of the previous three categories. We distinguish between pro-opposition and pro-Kiev accounts because a large portion of our data was collected during a period of active Russian involvement in the crisis in Ukraine. As a result, a discernible portion of Twitter accounts featured content critical of only one aspect of the Russian government activities: those pertaining to Crimea and Eastern Ukraine. Many of those accounts also carried easily

Figure 2. Distribution of orientation in the labeled set (1,946 bots).

identifiable features linking them to the Ukrainian cause, such as the country's flag or coat of arms. Another group of accounts featured a much broader array of anti-Kremlin content (issues related to Russian involvement in Ukraine were sometimes also present but did not dominate).

For the pro-Kremlin, pro-opposition, and pro-Kiev orientations, our coding instructions were very conservative. We wanted only accounts with a very explicit propaganda tone to them, featuring highly charged opinions *pro* or *contra*, to be assigned to these categories. However, similar to classification by type, in our instructions we listed easily identifiable characteristics of accounts in each group to simplify the task for coders and ensure consistency (see details in Supplemental Appendices C and E). For example, accounts aggressively targeting the Russian opposition and/or promoting Putin, his party, and ministers were labeled as pro-Kremlin. As a result of using a relatively narrow definition of partisanship, our neutral category is very broad by design and includes accounts with no, mixed,[6] or subtle ideological orientation, as well as an orientation that differs from the ones we focus on in this article.

Four coders coded each Twitter snapshot. We then aggregated the results with a weighted majority rule using the coders' trust score for weights and used accounts with intercoder reliability score above 0.75 in the labeled set (1,946 out of 2,413 bots, that is, 81%).

Figure 2 presents the distribution of orientation labels in the training, development, and test sets. As is evident, although pro-Kremlin bots were the most frequent category, other bot groups were prominently represented in the labeled sample.

To train a DNN for orientation detection, we used a number of textual features, including unigrams (words), bigrams (word pairs), hashtags, mentions, and links. On the one hand, we aimed to construct a large feature space that would make it possible to disentangle different types of political orientation; on the other, we wanted to avoid overfitting to the training set. With these two goals in mind, we retained only those features that appeared in tweets for at least 50 bots in the training set. Table 1 summarizes the distribution of the resulting 30,106 textual features. To make feature values comparable across accounts with different numbers of tweets, we normalized raw feature frequencies by the total number of tweets from a given account in our collection.

As Table 1 reveals, the data were relatively balanced across orientation groups and the training/development/test subsets.

To be clear, our approach relies on generating "ground truth" through human coding. The process described in this section—including training sessions with coders and only including accounts in the training data set with high intercoder reliability—is designed to make sure the accounts coded as bots do indeed match our understanding of what a

Table 1. Textual Data in the Labeled Set.

	Sets		
	Training	Development	Test
	(*N* = 1,557)	(*N* = 195)	(*N* = 194)
Unigrams and bigrams			
Total	27, 980	27,970	27,954
Pro-Kremlin	26,949	25,267	24,548
Neutral	24,862	18,267	16,878
Pro-opposition	27,311	23,233	23,510
Pro-Kiev	25,827	16,841	17,081
Hashtags			
Total	483	483	483
Pro-Kremlin	462	431	404
Neutral	369	104	177
Pro-opposition	469	387	371
Pro-Kiev	442	335	366
Links			
Total	333	333	333
Pro-Kremlin	310	279	258
Neutral	227	73	34
Pro-opposition	326	286	274
Pro-Kiev	280	218	224
Mentions			
Total	1,310	1,310	1,310
Pro-Kremlin	1,124	908	834
Neutral	24	161	75
Pro-opposition	1,267	1,096	981
Pro-Kiev	983	777	746

Note. Entries are frequencies of different sets of features (unigrams and bigrams, hashtags, URLs, and @ mentions) in the training, development, and test sets in total and by political orientation. Total number of features is 30,106. Total number of labeled accounts is 1,946.

bot is. There are, of course, costs to this type of approach: not having built the bots ourselves, at the end of the day we cannot know for sure that our accounts that look and act like bots are truly bots.[7] We are also relying on our own expectations of how bots appear and what they do to design our coding rules, and thus, it is possible that we could miss a new type of bot if we were not aware of its existence. Moreover, it is possible that an account that produced pro-opposition content was actually controlled by the Kremlin, or vice versa; our methods would not allow us to know whether this was the case.

Nevertheless, we believe our use of human coding to develop our training data set has a number of important advantages as well. First, the method is reproducible. The fact that we use saved tweets for our training data—as opposed to interacting with live Twitter—means that another group of researchers could show another group of coders the exact same tweets and give them the exact same coding instructions. Also valuable is the fact that this method can be used retrospectively with any existing collection of social media data, again because it does not require involving the Twitter API in the coding decision. Finally, generating ground truth through human-coded data expands the number of topics that researchers can study to anything that has generated enough data for human beings to code. Relying on leaked accounts, on the other hand, puts researchers at the mercy of what leakers choose to put into the public domain.

Table 2. Orientation Classifier Performance.

	Precision	Recall
Pro-Kremlin	0.97	0.99
Neutral	0.92	0.98
Pro-opposition	1.0	0.91
Pro-Kiev	0.97	1.0

Note. Entries are performance metrics for the test set.
Precision = $Pr(A \mid \hat{A})$ and Recall = $Pr(\hat{A} \mid A)$, where A is a given category of bots, and $\hat{A}$ is category A predicted.

Results

The MLP classifier with the best performance on the development set is a three-layer perceptron containing two hidden layers with 10 and 3 ReLUs and 0.0001 L_2 penalty hyperparameter.[8] Model weights are $10 \times 30{,}106$, 3×10, and 4×3 matrices. Coupled with the bias terms, the total number of estimated parameters is $301{,}119$.

The classifier shows high precision and recall on the test set, as Table 2 reveals (further details are available in Table 6 in Supplemental Appendix A).

When applied to the unlabeled set of 8,394 predicted bots with at least 10 tweets each year from 2015 to 2017, the MLP model makes high-confidence predictions for most observations, as Figure 9 in Supplemental Appendix A demonstartes. However, as the left panel of Figure 3 shows, the predicted orientation distribution is quite different from the randomly sampled labeled set, with a spike for the neutral category. Our further investigation of the driving forces of this spike revealed that the model overpredicted the neutral category for accounts with a small number of tweets in the collection. However, the problem disappeared when we focused on a subset of accounts with at least 100 tweets. One of the sources of the problem was that most accounts in the labeled set had more than 100 tweets (the lower quartile of the number of tweets in the labeled set is 140.5, whereas the upper quartile is 715). Thus, imposing the aforementioned restriction on the unlabeled set made it more similar to the data that the classifier used to learn its parameters. Besides, with fewer tweets, the MLP lacked data to reliably assign an account to one of the politically oriented groups, which is why the most probable category turned out to be the neutral one.

The right panel of Figure 3 shows the results of applying the MLP model as described above to bots with at least 100

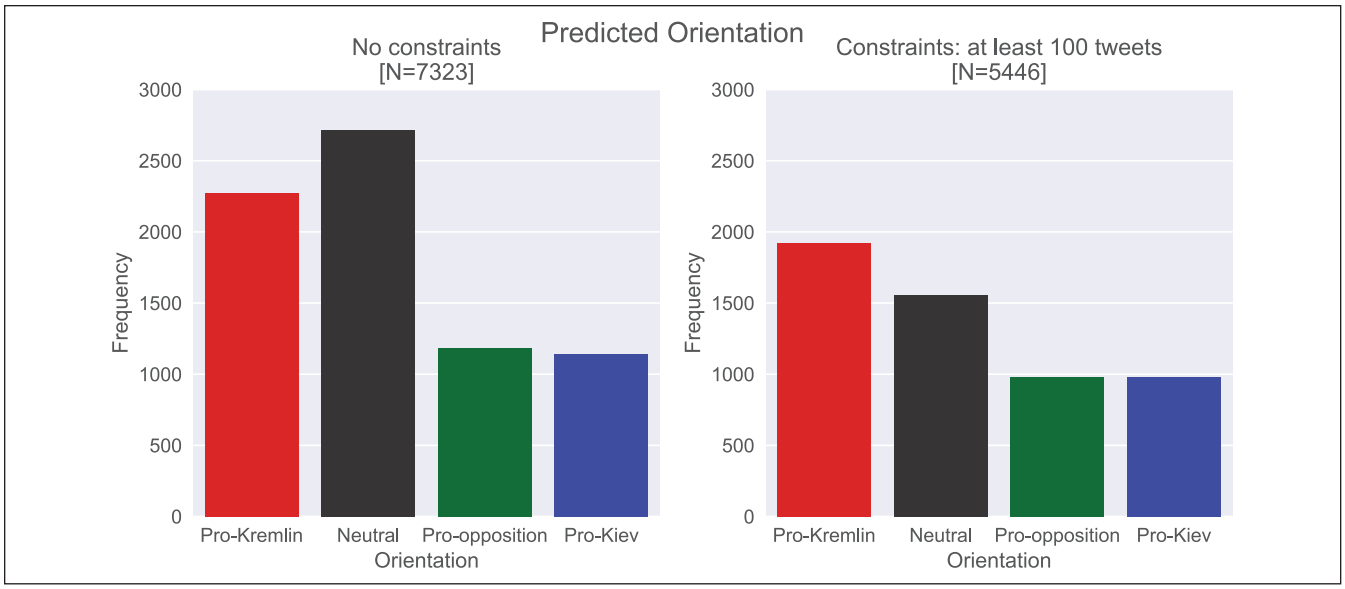

Figure 3. Predicted orientation distribution: Comparison.

tweets in our collection. As the figure suggests, the distribution of bots was as follows: 35% were pro-Kremlin, 18% were pro-opposition, 18% were pro-Kiev, and the remaining 29% were neutral.

However, different bots might be doing different things, and the number of bots is not necessarily the most important characteristic of bot presence on Twitter. Although a detailed analysis of bot activity is beyond the scope of this article, Figure 4 shows evidence that pro-Kremlin bots were much more active than neutral or anti-Kremlin bots in tweeting throughout the period under study. Thus, even though the number of anti-Kremlin bots was as high as the number of pro-regime ones, the latter had a larger presence in Russian political Twitter from 2015-2017.

Another interesting peculiarity of Russian Twitter bots is the gap between the average number of followers of neutral and politically oriented bots. Not only did the neutral bots have many more followers than pro- and anti-regime bots over all, but this gap grew over time. This was due to a constant increase in the number of accounts following neutral bots, whereas the number of followers of other types of bots was quite stable.

Thus, even though these results confirm that the Kremlin's narrative has been actively spread on Twitter by automated accounts, they also indicate that bots have been used to spread countervailing narratives on Twitter, from at least two different perspectives.

Verification and Further Exploration

In addition to evaluating model performance on the test set, we also attempted to verify the performance of the model in a number of other ways.

Our first test was to examine the model's out-of-sample predictive capacity using new human coding. To do so, we took a random sample of 244 bots from the restricted set of 5,446 accounts with at least 100 tweets in our collection and had each account coded by three of our original coders.[9] Six accounts produced large disagreements among coders and were not used in further analysis. We focus here on two subsets of results.

First, we consider the 238 accounts with intercoder agreement above 0.6 (i.e., at least two coders agreed on the label) and compare human codes with model predictions. The corresponding confusion matrix is presented in Table 3 and shows that the model does an excellent job of predicting out of sample.

These results still hold when we limit the analysis to the 162 accounts with complete coder agreement (shown in Table 4). Both tables show very good model performance on the unlabeled set.

As a second verification test, we examine how accounts with different political orientations differed in terms of the content they promoted most actively in their tweets. The idea here was to confirm that pro-Kremlin bots used in fact tweeting material that we might think of as pro-Kremlin and not, by contrast, pro-opposition or pro-Kiev. Figures 5 to 8, therefore, feature clouds with the most popular hyperlinks (URLs), mentions,[10] hashtags, and unigrams, respectively, by each orientation category of bots (pro-Kremlin, neutral, pro-opposition, and pro-Kiev), *as classified by our algorithm*. The color-coding, on the contrary, reflects *our assessment* of the relevant ideological association of that particular link/account/hashtag as pro-Kremlin (red), pro-opposition (green) or pro-Kiev (blue). Thus, the expectation is to see mostly red in the pro-Kremlin category, mostly green in the pro-opposition category, and mostly blue in the pro-Kiev categories.

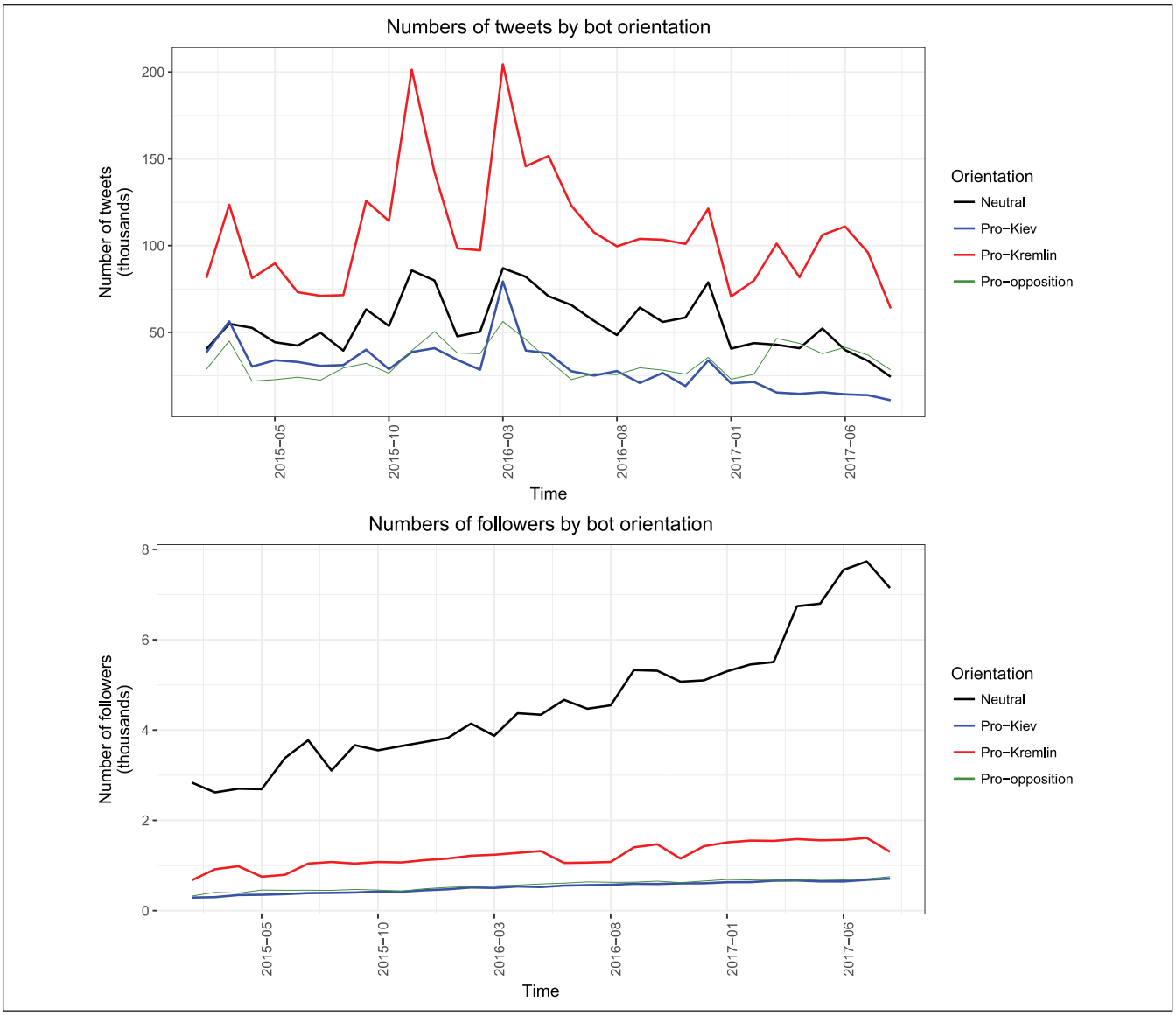

Figure 4. Predicted orientation distribution: Comparison.

Table 3. Verification: Confusion Matrix for Accounts With Coder Agreement 0.6.

	Predicted orientation			
	Pro-Kremlin	Neutral	Pro-opposition	Pro-Kiev
Coders' judgment				
Pro-Kremlin	82	2	0	0
Neutral	0	68	1	0
Pro-opposition	1	0	42	1
Pro-Kiev	0	0	1	40

Table 4. Verification: Confusion Matrix for Accounts With Coder Agreement 1.

	Predicted orientation			
	Pro-Kremlin	Neutral	Pro-opposition	Pro-Kiev
Coders' judgment				
Pro-Kremlin	59	2	0	0
Neutral	0	30	0	0
Pro-opposition	1	0	36	0
Pro-Kiev	0	0	0	34

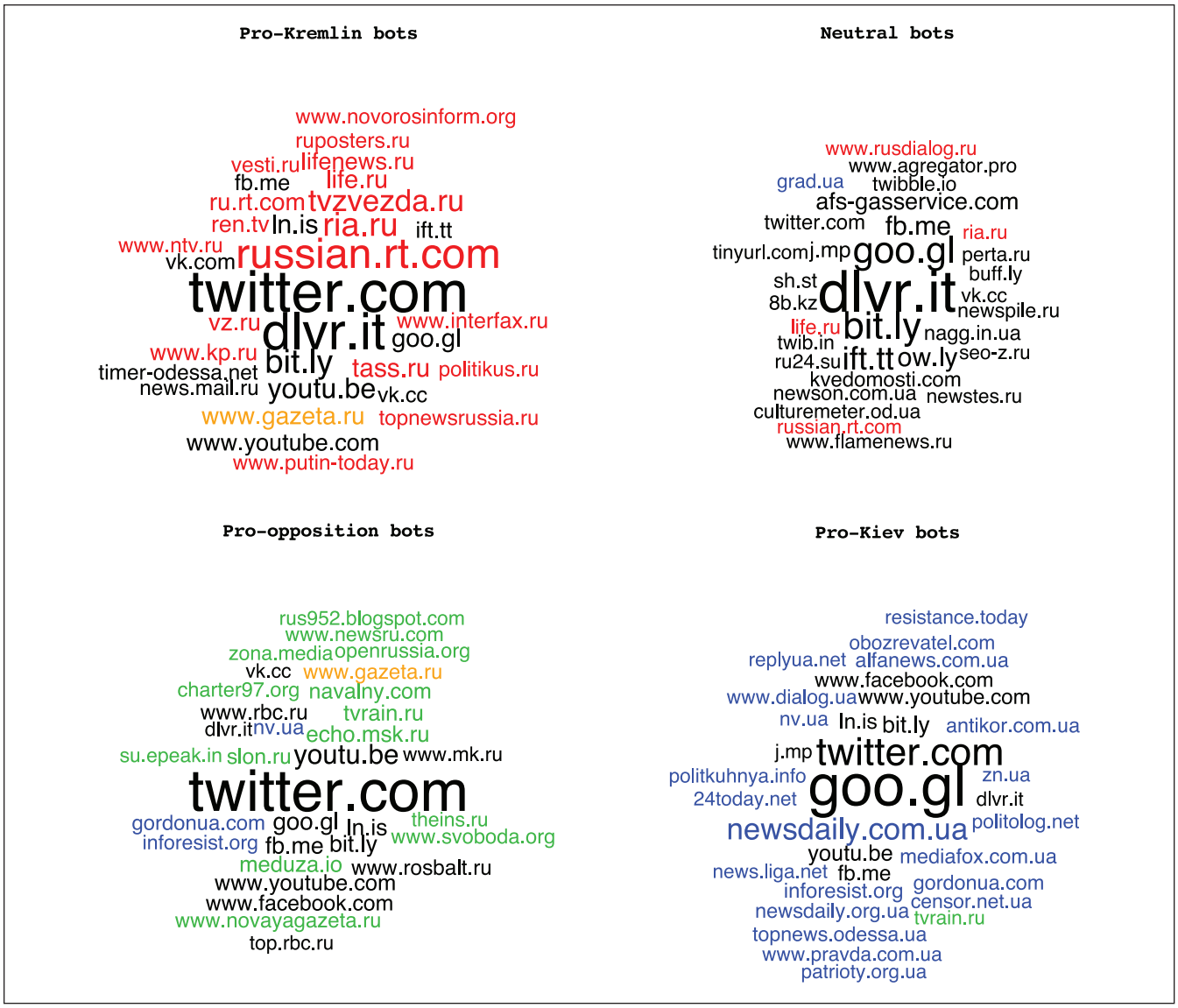

Figure 5. Top 32 URLs that appeared most frequently in tweets by each group of predicted bots (predicted probability $\geq .9$).
Note. Pro-Kremlin websites in red, pro-opposition in green, and pro-Kiev in blue; content-sharing platforms, news aggregators, and politically neutral websites are not color-coded (black). Gazeta.ru, which switched from being pro-opposition to pro-regime, is coded in orange. See Note 10 for details and caveats of color-coding.

For the purpose of determining the most popular hyperlinks tweeted by bots, we aggregate to the domain level (e.g., twitter.com). Figure 5 shows that all of the categories are heavily dominated by content-sharing platforms and news aggregators and, to a lesser degree, by partisan blogs and media. Importantly, though, while platforms such as Facebook, VKontakte, or YouTube are shared between bots of all orientations, media links are not.[11] Indeed, with few exceptions, pro-Kremlin, pro-opposition, and pro-Kiev bots include links to only those outlets that provide at least somewhat friendly news coverage.[12] This tendency is most pronounced for the pro-Kremlin bots, as every single media source on the list but one is either directly owned by the Russian government or is tightly controlled by entities associated with it. The list of top links for pro-opposition bots includes a couple outlets that could be considered neutral (*rbc.ru* and *rosbalt.ru*), but otherwise follows the same pattern, as does the list for pro-Kiev bots. The list of top links for bots that we classified as neutral includes mostly news aggregators and content-sharing platforms, as well as occasional media sources that are shared with pro-Kremlin and pro-Kiev bots.

At the same time, a close examination of the list reveals important differences. Among top six websites that most frequently appear in tweets by pro-Kremlin bots, three belong to the media: state-owned news agency *RIA* (a popular Russian news website and a parent company of both *RT* and *Sputnik*), the Russian version of *RT*, and *Zvezda*

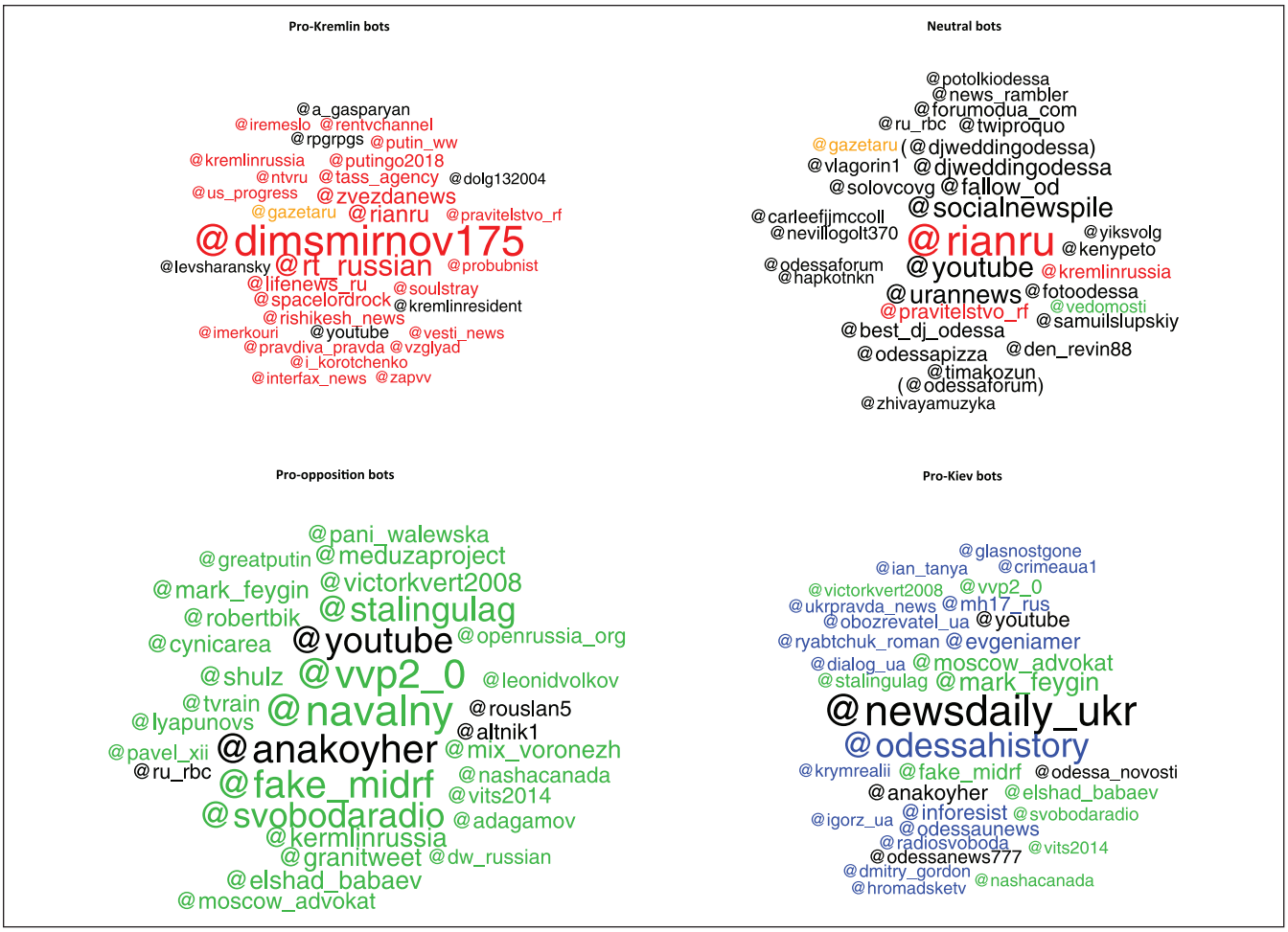

Figure 6. Top 32 mentions that appeared most frequently in tweets by each group of predicted bots (predicted probability ≥.9).
Note. Pro-Kremlin accounts in red, pro-opposition in green, and pro-Kiev in blue; accounts which do not clearly belong to either category are not color-coded. Gazeta.ru, which switched from being pro-opposition to pro-regime, is coded in orange. See Note 10 for details and caveats of color-coding.

(a military TV channel). Among the top six links for both pro-opposition and pro-Kiev bots, only one belongs to a media outlet, and the others are content-sharing platforms. Notably, all of those are foreign and include Facebook and YouTube. This shows the importance of content-sharing platforms that are not controlled by the Russian government for online adversaries of the Kremlin.

A particularly interesting case is *Gazeta.ru* (orange-colored in Figure 5). This online-only newspaper founded by late Russian Internet pioneer Anton Nossik was one of the most popular and respected sources of news in the Russian segment of the Internet. While maintaining editorial independence, it provided coverage of the Russian domestic opposition that could be characterized as either neutral or friendly. In recent years, the owner (through a number of successive editors) moved it much more to the pro-Kremlin side. However, this process was gradual compared with the abrupt dismissal of the editorial team at a fellow independent online news website *Lenta.ru* (Sanovich et al., 2018). The fact that *Gazeta.ru* is present on the lists of the most popular links for both pro-Kremlin and pro-opposition bots (at a higher position in the former) illustrates this dynamic.

The cloud of the most frequent Twitter accounts mentioned by bots in their tweets demonstrates an even higher degree of ideological consistency (Figure 6).[13] The separation between the lists for the pro-Kremlin and both anti-Kremlin categories is, again, perfect, even though the intersection between the pro-opposition and pro-Kiev lists is somewhat wider.

While the separation of links by the orientation of bots that tweet them is not particularly surprising (although the degree of the separation is remarkable nonetheless), the separation of mentions looks puzzling if the assumption is that one of their primary goals is to challenge arguments of the other side or, at the very least, simply clutter their mentions. If arguing is simply a domain of more sophisticated accounts (e.g., human trolls or cyborgs), even the most simple of bots can efficiently clutter mentions of accounts at whom they tweet. That we do not see this behavior in our data suggests

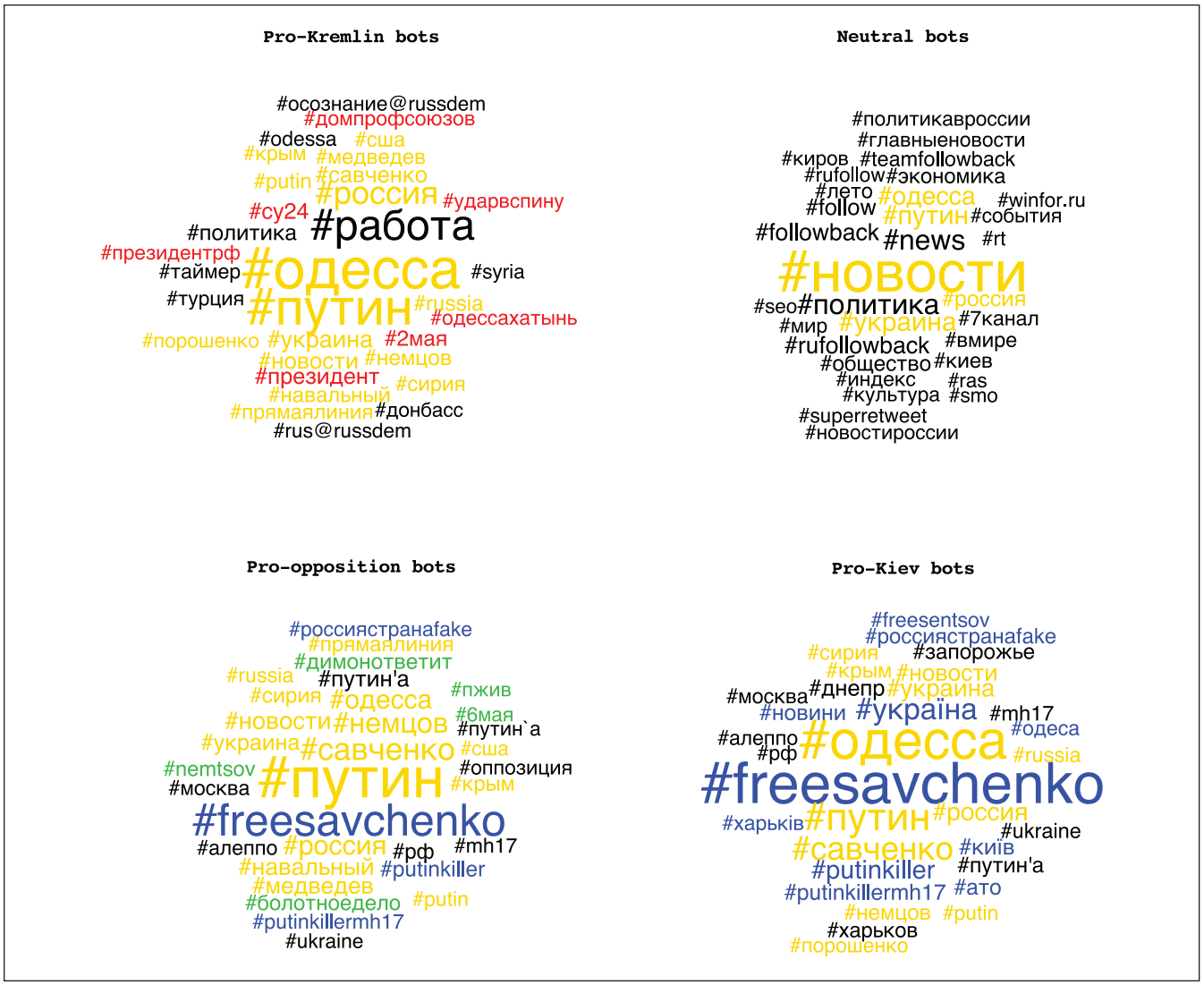

Figure 7. Top 32 hashtags that appeared most frequently in tweets by each group of predicted bots (predicted probability ≥.9).
Note. Hashtags that pro-Kremlin bots shared with either pro-opposition or pro-Kiev bots are in yellow. Among the rest: pro-Kremlin hashtags in red, pro-opposition in green, and pro-Kiev in blue; hashtags which do not clearly belong to either category are not color-coded. See Note 10 for details and caveats of color-coding.

that we might need to reevaluate our assumptions regarding the tasks that bots perform online.

Both the ideological consistency within groups of bots with the same political orientation and the separation between them appear to be highly robust to changes in the machine learning algorithm thresholds and aggregation rules we use for verification. Figure 10 (in Supplemental Appendix A) shows web links that appeared in the largest number of different bots from each group (as opposed to the total number of shares by all bots in a group), and when the threshold for predicted probability that a bot belongs to a group lowered from 90% to 70%. Figure 11 (also in Supplemental Appendix A) shows the cloud of Twitter accounts mentioned by the largest number of different bots from each group (as opposed to the total number of mentions by all bots in a group, again with the predicted probability for belonging to that political orientation kept at 70%). In both cases, the changes are inconsequential: Content consistency suffers for neutral bots only.

Compared with links and mentions, the most popular hashtags and words are naturally more generic and hence are used by bots in each group. In Figures 7 and 8, we highlight in yellow the most popular words and hashtags that were shared by pro-Kremlin with either pro-opposition or pro-Kiev bots. As Figure 7 reveals, many hashtags were indeed common for all three of these groups (and for neutral accounts too). Among hashtags that were not shared, there were a number of generic news-related hashtags, such as #odessa or #ukraine, in the original English. But in addition, in each group one can find hashtags that are clearly not generic, and, importantly for

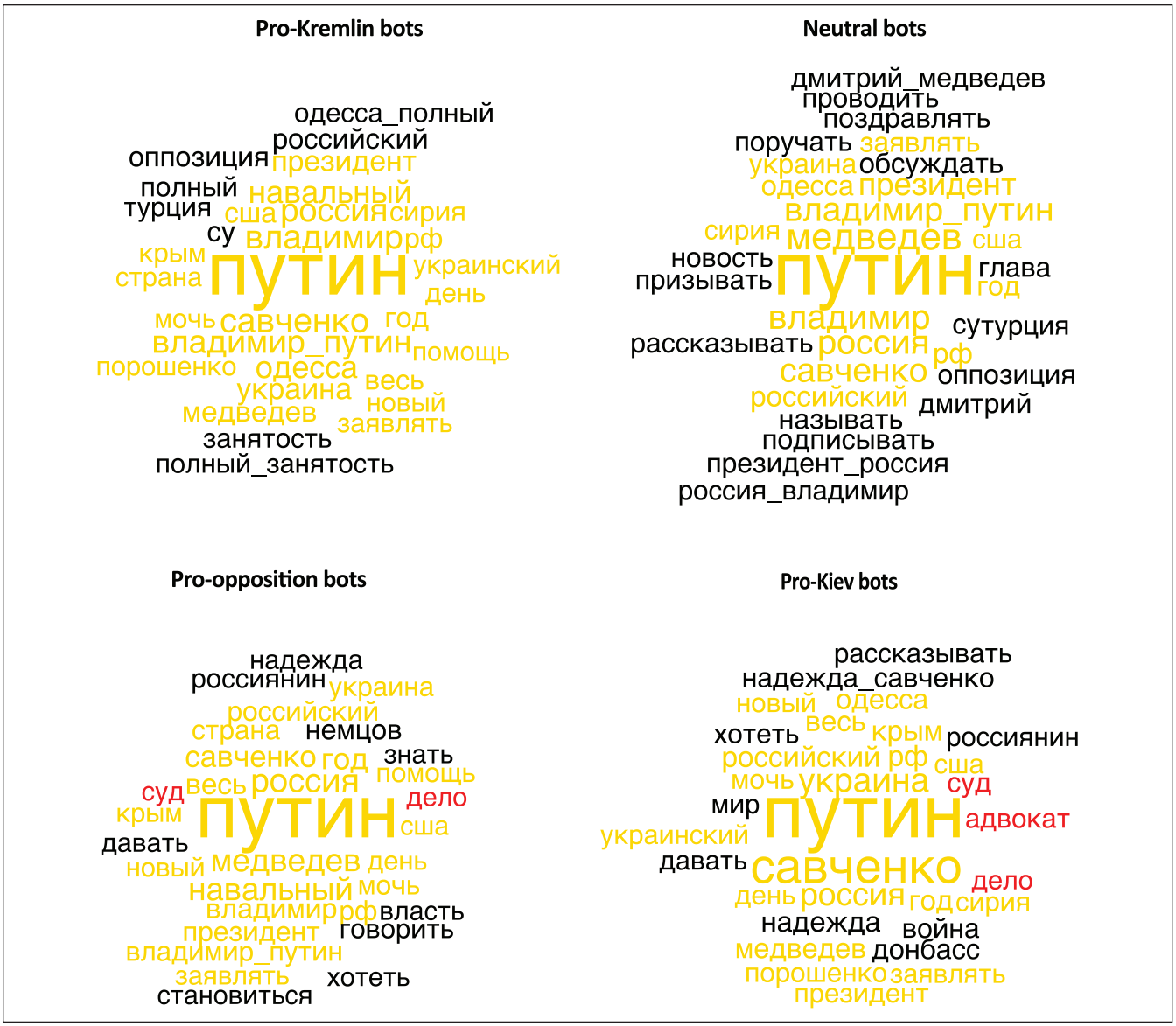

Figure 8. Top 32 words (unigrams) that appeared most frequently in tweets by each group of predicted bots (predicted probability ≥.9).
Note. Words that pro-Kremlin bots shared with either pro-opposition or pro-Kiev bots are in yellow. Words related to trial courts are in red. See Note 10 for details and caveats of color-coding.

verification purposes, they again perfectly match the orientation of the bots tweeting them. For instance, pro-Kremlin bots tweeted hashtags related to the tragic fire in Odessa on May 2, 2014, which the Kremlin blamed on Ukrainian nationalists. Conversely, pro-Kiev bots tweeted hashtags that directly blamed Putin for shooting down the MH17 plane on July 17, 2014, near Donetsk (Golovchenko, Hartmann, & Adler-Nissen, 2018). The top hashtags of pro-opposition bots refer to a big court case against opposition activists after a rally at Bolotnaya Square in Moscow and to a campaign to demand that the Russian Prime Minister Dmitry Medvedev respond to corruption accusations. While the words used by bots (Figure 8) are even less group-specific, the notable exception is that both pro-opposition and pro-Kiev bots (but not pro-Kremlin ones) featured words related to legal proceedings ("court," "case," and, for pro-Kiev bots, "attorney"). Both opposition activists and Ukrainian soldiers and activists were on trial in Russia during this period.

While examining the most popular words and hashtags that are largely shared across groups of bots could be less illuminating than contrasting clearly separable links and mentions, it is the comparison that might be most instructive here. While pro-Kremlin bots do not tweet at Alexei Navalny, a major Russian opposition leader, and neither pro-Kiev nor pro-opposition bots post links to RT, the pro-Russian government news source,

pro-Kremlin bots do use the hashtag #Navalny and both pro-opposition and pro-Kiev bots often mention Putin and Medvedev. In other words, bots do follow the laws of political advertising: paint your opponent, but do not let them speak.

Conclusion

Previous research has suggested that the Russian political Twittersphere may be swamped with bots. However, precisely due to its large scale, the nature of this activity is hard to dissect in a systematic, reliable, and reproducible way. One option for conducting this type of research would be to rely on externally determined attributions of accounts (e.g., a leak, a whistleblower, accounts identified by Twitter and made public.). While valuable, this type of approach also has limitations. First of all, if the list of controlled bots is made public, they usually stop their activity, in many cases because the accounts are shut down by Twitter. Second, this approach, while suitable for demonstrating examples of politically charged bot activity, is not well suited to estimating of its prevalence in a larger network of interest. Finally, given that leaks are infrequent and often take place much later than the bulk of bot activity, such an approach is clearly not a reliable foundation for ongoing research, particularly when temporal comparisons are of interest.

Thus, an important contribution of our study is the presentation of a systematic, retrospective, and reproducible approach to identifying bots' political affiliations that is not based on externally supplied labels for bots and demonstrates robust results in a number of specifications. In particular, we present a neural network model that uses textual data (unigrams and bigrams, hashtags, mentions, and links) in the bag-of-words framework to predict the political orientation of bots that tweeted about Russian politics in 2015-2017. We train an MLP with two hidden layers of ReLUs on a relatively small training set of about 1,500 observations and achieve high precision and recall (above 90%) for pro-regime, anti-regime, and neutral bots.

We also apply our model to an unlabeled set of more than 6,000 accounts (predicted to be bots by our bot detection model) and show that the political orientation model performs well when it is applied to a subset of accounts with enough tweets. However, we find that the model overpredicts neutral accounts when applied to bots with less than 100 tweets in our collection. From a more substantive perspective, we show that—contrary to common wisdom—pro-Kremlin bots are not the only type of bots in Russian political Twittersphere. Indeed, anti-Kremlin bots of two different types also maintained a non-trivial presence on Russian Twitter. In addition, we find a large group of neutral bots that we believe could be used by mass media for search engine optimization purposes. This suggests that a systematic analysis of the use of Twitter bots in Russia could not be implemented without distinguishing between automated accounts with different political affiliations.

Both pro- and anti-Kremlin bots exhibited a high degree of consistency in promoting only those top accounts on Twitter that were clearly on their side of the political spectrum. Their interaction with media, however, followed different patterns: pro-Kremlin bots utilized news stories available at Kremlin-controlled websites; anti-Kremlin bots relied upon content hosted by platforms that the Kremlin could not control.

More generally, our findings highlight the fact that any attempt to characterize the political activity of bots is always going to be a two-step process. Methods to detect bots—while of course necessary for any such enterprise—are not sufficient: researchers must also take the additional step of classifying bots by political orientation. Otherwise, we face a real risk of attributing political motivations to bots that at best may be neutral and at worst may actually be supporting the opposite side in a political conflict. Fortunately, our work also shows that building tools to classify bots by political orientation is not only necessary but also possible. Thus, our hope is that the methods contained in this article will enable researchers to carry out a more systematic and large-scale analysis of the use of computational propaganda tools, as well as provide the foundations for a more detailed and nuanced study of the use of Twitter bots in Russia.

Authors' Note

Sergey Sanovich is also affiliated to Stanford University, Center for International Security and Cooperation, Palo Alto, CA, USA. Richard Bonneau is also affiliated to Flatiron Institute, Center for Computational Biology, New York, NY, USA. Stukal, Sanovich, Tucker, and Bonneau designed the study. Stukal developed the code for all of the analysis and visualization. Stukal prepared the first draft of the manuscript, except the Verification and Further Exploration section. Sanovich prepared the first draft of the Verification and Further Exploration section and Appendices C - F used for building the training set. Stukal recruited and Sanovich trained human coders. Both Sanovich and Stukal supervised human coders. Bonneau and Tucker oversaw the data collection process. All of the authors participated in revising and editing of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors gratefully acknowledge the financial support for the NYU Social Media and Political Participation (SMaPP) lab from the National Science Foundation (Awards SES-1248077; SES 1756657), the William and Flora Hewlett Foundation, the Rita Allen Foundation, the Knight Foundation, the Bill and Melinda Gates Foundation, Craig Newmark Philanthropies, the Democracy Fund, the Intel Corporation, the New York University Global Institute for Advanced Study, and Dean Thomas Carew's Research Investment Fund at New York University.

Supplemental Material

Supplemental material for this article is available online at: https://smappnyu.org/wp-content/uploads/2019/03/Supplemental_Appendix.pdf.

Notes

1. The analysis focused on the accounts that tweeted at least 10 times about Russian politics, as defined by our collection of keywords.
2. The reason each classifier is run on 10 different training sets is because the collection of hand-coded Twitter accounts was unbalanced. To cope with the pitfalls of analyzing highly imbalanced data sets, the authors produced fully balanced training sets by using all available human accounts and taking a subsample of labeled bots of the same size as the number of labeled humans. As this approach, however, could bring additional stochasticity into the classification results, the process was repeated 10 times on 10 different training sets with identical labeled human accounts, but different randomly chosen bot examples. The ensemble yields the test set precision of 0.99 and recall of 0.77, thus making us very confident that the accounts that were labeled as bots are indeed automated accounts.
3. Another consequence of this complexity is the difficulty with using pretrained word embeddings (Kutuzov & Kuzmenko, 2017), as different bots could use the same words for different purposes and in different contexts. Our experiments with pretrained word embeddings produced much weaker results than the bag-of-words approach that we adopt in this article.
4. The first number indicates the number of units in the first hidden layer, and the number after the hyphen refers to the second hidden layer.
5. We created Twitter snapshots using a special Python module that is publicly available at: https://github.com/denisStukal/twitter_bots.
6. For example, an account that praises the Kremlin's foreign policy but loathes its economic decisions.
7. Of course, the only way we could have truly known whether accounts were indeed bots would be if we had built them ourselves (or paid someone to build them). The problem in that case, however, would be that our ultimate goal is to characterize the behavior of bots in the Russian political Twittersphere. If we programmed the bots ourselves, we would not learn anything of use from observing their behavior.
8. As suggested by an anonymous reviewer, we also considered using dropout by Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2017) as an alternative regularization strategy. Model performance with dropout shows very similar performance and is reported in Table 5 in Supplemental Appendix A.
9. Ideally, we would have had each account coded by four coders as we did in the first round of coding. However, when we reached out to the same group of human coders and asked them to do the additional work for extra payment, only four out of seven people agreed. Given the smaller number of available coders, we elected to have each account coded by three coders.
10. Mentions are direct references to other Twitter user in Tweets by referencing that Twitter user's "handle" or account name.
11. The only exception to this rule is not surprising: Lists for pro-opposition and pro-Kiev bots share a few websites (*Gordonua.com, Inforesist.org, nv.ua*, and *TVRain.ru*) between themselves, but nothing with the pro-Kremlin list.
12. How friendly they are is not the subject of our research. Our color-coding serves only the purposes of identifying *trends within groups of bots* and is *not* intended to characterize any media resource or blog in particular. We further note that media sources in these lists vary dramatically both in the degree of ideological affinity (from independent media such as *Meduza.io* that merely cover the opposition in a fair way to an opposition leader's campaign website *Navalny.com*) and in the quality of news coverage (from a respected albeit state-owned news agency *Interfax.ru* to a pro-Kremlin tabloid *Lifenews.ru*). The only aspect of our color-coding that is important for the validation is that colors could not be plausibly *switched* between pro-Kremlin and pro-opposition/pro-Kiev websites. As the patterns are very consistent, adding or removing the color code for a few items would not change our assessment of the verification.
13. Some Twitter accounts in this list (as well as a few URLs in Figure 5) are no longer available or were hijacked and no longer belong to their original owners. In these cases, ideological affiliation was reconstructed based on media reports, Internet archives, and other sources.

References

Alandete, D. (2017, November 11). Russian network used Venezuelan accounts to deepen Catalan crisis: An analysis of five million messages reveals that RT and Sputnik used social networks to spread a negative image of Spain. *El País*. Retrieved from https://elpais.com/elpais/2017/11/11/inenglish/1510395422_468026.html

Arnsdorf, I. (2017, August 23). Pro-Russian bots take up the right-wing cause after Charlottesville: Analysts tracking Russian influence operations find a feedback loop between Kremlin propaganda and far-right memes. *ProPublica*. Retrieved from https://www.propublica.org/article/pro-russian-bots-take-up-the-right-wing-cause-after-charlottesville

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, *21*(11). Retrieved from https://firstmonday.org/article/view/7090/5653

Booth, R., Weaver, M., Hern, A., & Walker, S. (2017, November 14). Russia used hundreds of fake accounts to tweet about Brexit, data shows. *The Guardian*. Retrieved from https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets

Brachten, F., Stieglitz, S., Hofeditz, L., Kloppenborg, K., & Reimann, A. (2017). Strategies and influence of social bots in a 2017 German state election—A case study on Twitter. arXiv:1710.07562 [cs]. arXiv:1710.07562.

Brenner, S. (2007). "At light speed": Attribution and response to cybercrime/terrorism/warfare. *The Journal of Criminal Law and Criminology*, *97*, 379-475.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*, 811-824.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017, April 3-7). 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 (pp. 963-972). arXiv:1701.03017

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*, 303-314.

Dietterich, T. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Proceedings of the first international workshop on multiple classifier systems 2000, lecture notes in computer science 1857* (pp. 1-15). Berlin, Heidelberg: Springer Verlag.

Entous, A., Nakashima, E., & Jaffe, G. (2017, December 25). Kremlin trolls burned across the Internet as Washington debated options. *Veterans Today*. Retrieved from https://www.veteranstoday.com/2017/12/25/kremlin-trolls-burned-across-the-internet-as-washington-debated-options/

Forelle, M., Howard, P., Monroy-Hernndez, A., & Savage, S. (2015). Political bots and the manipulation of public opinion in Venezuela. arXiv:1507.07109 [physics]. arXiv:1507.07109

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017). An in-depth characterisation of Bots and Humans on Twitter. arXiv:1704.01508 [cs]. arXiv:1704.01508

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of machine learning research* (pp. 315-323). Retrieved from http://proceedings.mlr.press/v15/glorot11a.html

Goldberg, Y. (2017). *Neural network methods in natural language processing*. San Rafael, CA: Morgan & Claypool.

Golovchenko, Y., Hartmann, M., & Adler-Nissen, R. (2018). State, media and civil society in the information warfare over Ukraine: Citizen curators of digital disinformation. *International Affairs*, *94*, 975-994.

Gorwa, R. (2017, March 20). On the Internet, nobody knows that you're a Russian bot. *Council on Foreign Relations*. Retrieved from https://www.cfr.org/blog/internet-nobody-knows-youre-russian-bot

Gorwa, R., & Guilbeault, D. (2018). Understanding bots for policy and research: Challenges, methods, and solutions. arXiv:1801.06863 [cs]. arXiv:1801.06863

Grassegger, H., & Krogerus, M. (2017, December 2). Fake news and botnets: How Russia weaponised the web. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2017/dec/02/fake-news-botnets-how-russia-weaponised-the-web-cyber-attack-estonia

Grimme, C., Preuss, M., Adam, L., & Trautmann, H. (2017). Social bots: Human-like by means of human control? *Big Data*, *5*, 279-293.

Hegelich, S., & Janetzko, D. (2016, March 31). *Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet*. Tenth International AAAI conference on Web and Social Media. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015

Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2017). Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580. arXiv:1207.0580.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359-366.

Keller, F., Schoch, D., Stier, S., & Yang, J. (2017, May 3). How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. In *International AAAI conference on Web and Social Media* (pp. 564-567). Retrieved from https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15638

King, G., Pan, J., & Roberts, M. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*, 484-501.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012, December 3-6). ImageNet classification with deep convolutional neural networks. In NIPS'12 proceedings of the 25th international conference on Neural Information Processing Systems (Vol. *1*, pp. 1097-1105). Retrieved from https://dl.acm.org/citation.cfm?id=2999257

Kutuzov, A., & Kuzmenko, E. (2017). Webvectors: A toolkit for building web interfaces for vector semantic models. In *Communications in Computer and Information Science. Analysis of images, social networks and texts 2016*. doi:10.1007/978-3-319-52920-2_15

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444.

Miller, B. (2017). *Automated detection of Chinese government astroturfers using network and social metadata*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2738325

Nimmo, B. (2017). *#ElectionWatch: Russia and referendums in Catalonia?* Retrieved from https://medium.com/dfrlab/electionwatch-russia-and-referendums-in-catalonia-192743efcd76

Oentaryo, R., Murdopo, A., Prasetyo, P., & Lim, E.-P. (2016). On profiling bots in social media. In *Social informatics* (Vol. *10046*, pp. 92-109). doi:10.1007/978-3-319-47880-7_6

O'Sullivan, D., & Byers, D. (2017, September 28). Fake black activist social media accounts linked to Russian government. *CNN*. Retrieved from https://money.cnn.com/2017/09/28/media/blacktivist-russia-facebook-twitter/index.html

Popken, B. (2017, November 30). Russian trolls' graphic tweets on racism, rape, and Satanism revealed. *NBC News*. Retrieved from https://www.nbcnews.com/tech/social-media/russian-trolls-pushed-graphic-racist-tweets-american-voters-n823001

Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Flammini, A., & Menczer, F. (2011, July 5). Detecting and tracking political abuse in social media. In Proceedings of the fifth international AAAI conference on Weblogs and Social Media (CWSM) (pp. 297-304). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850

Rosenberg, E. (2018, January 19). Twitter to tell 677,000 users they were had by the Russians. Some signs show the problem continues. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2018/01/19/twitter-to-tell-677000-users-they-were-had-by-the-russians-some-signs-show-the-problem-continues/?utm_term=.2d4437679411

Sanovich, S., Stukal, D., & Tucker, J. A. (2018). Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia. *Comparative Politics*, *50*, 435-482.

Schfer, F., Evert, S., & Heinrich, P. (2017). Japan's 2014 general election: Political bots, right-wing internet activism, and prime minister Shinz Abe's hidden nationalist agenda. *Big Data*, *5*, 294-309.

Shane, S. (2017, September 7). The fake Americans Russia created to influence the election. *The New York Times*. Retrieved

from https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html

Sobolev, A. (2018). *How pro-government "trolls" influence online conversations in Russia*. Retrieved from https://asobolev.com/files/Anton-Sobolev-Trolls.pdf; Unpublished manuscript.

Stewart, L., Arif, A., & Starbird, K. (2018). Examining trolls and polarization with a retweet network. In *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. Retrieved from https://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf

Stieglitz, S., Brachten, F., Berthel, D., Schlaus, M., Venetopoulou, C., & Veutgen, D. (2017). Do social bots (still) act different to humans? Comparing metrics of social bots with those of humans. In G. Meiselwitz (Ed.), *Lecture Notes in Computer Science. Social computing and social media. Human behavior* (pp. 379-395). Cham, Switzerland: Springer.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). Do social bots dream of electric sheep? *A categorisation of social media bot accounts*. arXiv:1710.04044 [cs]. arXiv:1710.04044

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting bots on Russian political Twitter. *Big Data*, *5*, 310-324.

Szegedy, C., Liu, W., Jia, Y. J., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015, June 7-12). Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1090-1098). Retrieved from https://ieeexplore.ieee.org/document/7298594

TASS (2017). *Peskov: Kremlin has never placed political ads on Facebook*. (In Russian). *TASS*. Retrieved from https://tass.ru/politika/4584093

Varol, O., Ferrara, E., Davis, C., Menczer, F. & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. arXiv:1703.03107 [cs]. arXiv:1703.03107

Wells, G., & Seetharaman, D. (2017, October 13). Facebook users were unwitting targets of Russia-backed scheme. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/facebook-users-were-unwitting-targets-of-russia-backed-scheme-1507918659

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. arXiv:1801.09288 [cs]. arXiv:1801.09288

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. Boca Raton, FL: Chapman & Hall/CRC Press.

Author Biographies

Denis Stukal is a postdoctoral fellow at the Social Media and Political Participation (SMaPP) Laboratory at New York University. He holds a PhD in Politics from NYU (2018). His research focuses on data science, computational communication, and political methodology. He is especially interested in analyzing large corpora of texts and collections of other social media data to study misinformation and computational propaganda. His research has been published in journals such as *Big Data*, the *Journal of Politics*, and *Comparative Politics*.

Sergey Sanovich is a cybersecurity postdoctoral fellow at Stanford University Center for International Security and Cooperation (CISAC). He received his PhD in Politics at NYU, working at the Social Media and Political Participation (SMaPP) Lab. He studies how autocrats use the power of persuasion to come to, and stay in, office. His current research is focused on online censorship and propaganda by authoritarian regimes; social media platform governance; elections and partisanship in electoral autocracies; and personalization of politics in bothautocratic and democratic countries.

Joshua A. Tucker is professor of politics, affiliated professor of Russian and Slavic Studies, and affiliated professor of data science at New York University. He is the director of NYU's Jordan Center for Advanced Study of Russia, a co-director of the NYU Social Media and Political Participation (SMaPP) laboratory, and a co-author/editor of the award-winning politics and policy blog The Monkey Cage at The Washington Post. His research has appeared in over two-dozen scholarly journals, and his most recent book is the co-authored *Communism's Shadow: Historical Legacies and Contemporary Political Attitudes* (Princeton University Press, 2017).

Richard Bonneau is professor of biology, computer science, and data science at New York University. He is a co-director of the NYU Social Media and Political Participation (SMaPP) laboratory. He is also group leader for Systems Biology at the Center for Computational Biology in Flatiron Institute, The Simons Foundation. His research focuses on creating new methods for protein structure design and modeling as well as new methods for understanding social and biological networks.