School of Informatics

Informatics Project Proposal
Expressing or concealing the self: the choice of online usernames and profile content for (im)personal identity

[redacted]

Abstract

Previous evidence suggests that a worryingly amount of social media content on Twitter, and not only, is being actively generated by automated entities, widely known as social bots. This project proposes a Machine Learning approach for detecting political bots, with a focus on the 2015 Brexit campaign in the UK. The aim of this project is to contribute to the ever expanding field of social network analysis and to further increase awareness on the effect of political bots on Twitter.

Date: Monday 27th April, 2020

[redacted]

1 Motivation

Social Networking Services (SNS) have become part of our everyday lives, connecting millions of people across the globe and facilitating information exchange as never seen before. Those powerful social tools help create a layer of connections that promotes information broadcast, which, in turn, affects the news and opinions we are exposed to. Since many people have turned to social media to read and disseminate news, various entities have started using this medium as a means to spread misinformation and skew the social, political and economic climate by promoting certain trends or public figures.

Bots – bits of computer code designed to operate an account and sometimes even masquerade as a real person by exhibiting human-like communication behavior – are shaping our daily online experience by altering, disrupting or silencing online conversations artificially. Twitter, a popular social networking service, is plagued by those types of automated accounts. One study from 2017 has estimated that the number of bots on Twitter lies somewhere between 9-15% [1], which means that as least 48 million accounts on Twitter are not human. Therefore, a deep reliance on social media may make us vulnerable to manipulation. Moreover, this manipulation also bleeds across into the analysis that is done on social networking platforms. When the content researchers are studying is filled with automated software programs pretending to be people, the quality of their work might suffer greatly and false conclusions might be drawn.

An important insight into the problem of bots is that they don't create new conflict. Bots always exploit existing conflicts, existing cracks in our political systems for example, and then exacerbate these cracks and drive wedges into them. For instance, bots were responsible for millions of tweets right before the 2016 election [2]. Furthermore, a high volume of tweets was generated a few days before the voting day of the 2016 Brexit Referendum, and then dropped afterwards [3] [4]. This was not the first time a disinformation campaign was coordinated through bots, with further examples including the 2017 French presidential election[5], the German Presidential Election[6] and the Syrian Civil War[7]. As (Ferrara et. al, 2016)[8] explained, *"The novel challenge brought by bots is the fact that they can give the false impression that some piece of information, regardless of its accuracy, is highly popular and endorsed by many, exerting an influence against which we haven't yet developed antibodies"*. Therefore, given their malicious behavior and potential real-life consequences their actions might have had in past events by possibly shifting public opinion, an effective and precise way of identifying bots is sorely needed, as to hopefully avoid the consequences their actions might have in the future.

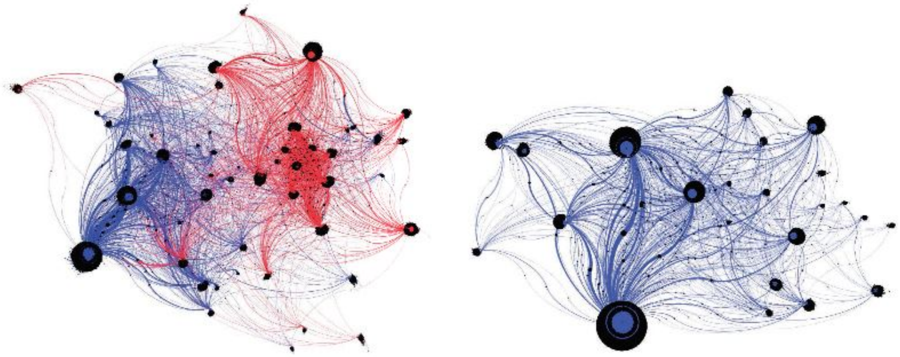

Figure 1: An graph of bot (red) and human (blue) interactions to highlight how influential bots are, as seen in [9].

2 Hypothesis

Our hypothesis is that there is a combination of name and behavior that constitutes suspicious or disingenuous agents (human of artificial), and this combination can be detected by employing different supervised Machine Learning techniques. If successful, this approach will serve for better identification of malicious bots on social media platforms.

The novelty of the approach doesn't come from the method used or from new-found features, but rather from the feature space selection. The features chosen were influenced by the works of others, as it will be discussed in the next section. We want to select the least number of features that will yield the best result, all the while aiming for a fast and easy to update or re-train classifier. Essentially, we are asking the following main question: what features are the most useful, the richest in information, for bot vs human discrimination? How much can we reduce the feature space while still having good results? Do we really need a plethora of features, or just a small, carefully selected number in order to achieve a good performance?

Additionally, we hope to bring some insight into the bot activity and influence during the Leave and Remain campaign during and after the Brexit Referendum by offering a thorough comparison between bots and humans. We will use the features collected for our bot vs human classifier and analyse them to build up a comprehensive understanding on bot vs. human behavioural characteristics, by comparing their glaring differences and striking similarities. We aim to answer several question, such as: what are the typical behaviors of humans and bots? What is the most obvious difference, or their most surprising similarity? Do bots or humans generate the most content? What about interaction: what is the main difference between bot-to-bot interaction vs human-to-human?

3 Background

In this section we will give a brief review on some of the most important background information regarding the project. Specifically, we will introduce and describe the social network from where all our data was collected: Twitter. Furthermore, we will also mention more about the dataset used, as well as its limitations.

3.1 Twitter

Twitter is a social networking and micro-blogging service, which was first launched in 2006. It quickly gained popularity, reaching more than 100 million users by 2012 and, as of 2018, boasting more than 321 million monthly active users. Twitter has a rather unique mechanism, where users follow other users. If you follow someone, you can see their tweets in your timeline. The platform allows users to interact in a variety of ways, such as:

- **Tweets**: Short messages (microblogs) that user can post. They used to have a length restriction of 140 characters, but as of November 2017 it was upped to 280. Tweets can contain text, URLs, images, and videos.
- **Retweets**: Users share information that has been already tweeted by others, without adding a comment. They appear on the user's followers' timeline.
- **Quotes**: The same principle behind a retweet, but now the user has also made a comment regarding the tweet they quoted, starting a new thread.

- **Replies**: A user replies to a tweet that someone else made.
- **Likes**: A user can like a tweet. Tweets become popular when they have a large number of retweets and/or likes.
- **Mentions**: A user can mention another user in a tweet, by putting the @ symbol before their username. That user will get a notification.
- **Hashtags**: Handles (#) that are used to aggregate tweets about the same subject.
- **Trends**: A list of hashtags and terms most popular on Twitter at any moment in time.
- **Followers**: People who follow a user. A follower is defined as a directed edge from user A to user B, such that user A will see what user B posts, but not the other way around
- **Friend**: Two users, A and B are friends if user A follows user B and user B follows user A back. Simply put, both users will see what each other posts if they are friends.

User activity vary significantly. Some users might not tweet at all, only consuming content, without actually engaging ("lurkers"). Other might use it very often, tweeting and retweeting many times per day. Some might log on rarely, giving the impression of an abandoned account, while others might log on multiple times per day. Many might fall somewhere in-between.

3.2 Dataset

For building the detection model, we will be using an existing dataset of partially annotated tweets involving the Leave and Remain campaigns during and after the Brexit Referendum. The dataset includes bots and trolls, as well as ordinary accounts, with a much larger percentage consisting of regular users. The data was collected using the Twitter API, based on a selection of relevant hashtags chosen by a panel of academic experts (examples of hashtags used: #eureferendum, #euref, #brexit, #no2eu, #yes2eu, #notoeu, #yestoeu, #betteroffout, etc.) [10][11]. The dataset contains over 70 million tweets, and the set of hashtags has been updated periodically since August 2015 to better reflect the ongoing conversation on Brexit.

4 Methodology

As it was previously mentioned, the success of this project and the quality of the analysis we wish to conduct on this dataset relies heavily on the choice of features and model. In this section we will give an overview of the methodology used in the project. Specifically, we will talk about the selection of the most appropriate models for the task at hand, as well as discuss the features selected and the main influences behind the decision to choose this particular feature space.

4.1 Model Selection

In the past years, bots have become increasingly more complex, exhibiting more human-like behavior in order to avoid detection. However, several advanced detection algorithms have appeared in order to keep up with this evolution. The algorithms range from very simple to very complex. We can place the current Twitter bot detection algorithms into two categories: supervised and unsupervised.

Unsupervised machine learning algorithms, that need to training data, have not been very popular methods in detecting bots, with only few notable papers using them. While unsupervised models are more likely to detect novel bot behavior, that might fool a supervised model trained only on a subgroup of bots, unsupervised models are harder to evaluate due to the absence of labeled data. The unsupervised models used usually include PCA-based [12], Stream-based [13] and correlated pairwise similarity [14]. A notable example in this line of research is considered DeBot [14], which uses dynamic time warping to identify bots with synchronized tweeting patterns.

At the other end, *supervised machine learning algorithms* have been and still are the most used methods for bot detection. Given a reasonably large and annotated dataset, a model can be trained to predict bots with an accuracy of, usually, over 90%. However, one obvious flaw in employing a supervised approach is the risk of overfitting (i.e. not being able to generalize well on unseen data). Since every supervised model is trained on a particular chunk of the twitter database, it might run the risk of not being able to detect bot behavior that was not present in the training data. The supervised models used include: Neural Networks [15], Naive Bayes [16], Logistic Regressions[17], SVM [18] and Random Forest. In this case, the current state of the art is, to our knowledge, considered to be BotOrNot (now Botometer) [19], which is a Random Forest classifier that uses 1,150 features to learn decision rules in order to distinguish between humans and bots.

For our model we will use a supervised approach. In most cases, literature recommends a Random Forest algorithm. However, given the variety of models used by others and their good results, we will also look into other algorithms, such as: XGBoost, Adaptive Boosted Decision Trees and Neural Networks.

4.2 Feature space selection

4.2.1 Profile features

Profile features, which are features that can be extracted from the user's profile information, are widely used in literature and have been proven to yield good results [20] [21]. The most commonly used features are given by numerical characteristics, such as: number of followers, number of followings, statuses count and various ratios that can be obtained by combining them (e.g. followers-to-friends, which is probably the most popular ratio used) [22] [23].

Several other profile features can be harnessed from user meta-data, such as time since creation, which has been proven to be a useful feature given the research of [24] on Russian Trolls and of [25] on the Russian parliamentary election of 2011. Furthermore, it has been hypothesized that, since human effort and imagination is required to customize one's profile (by changing, for example, the profile description, background image or profile image) [26], a bare profile or a barely customized one can be a signal that the account is probably a bot, so checking those details can be helpful [27] [1] [23].

Additionally, [28] showed that simply the profile description and name can be indicators of a non-human account. Even when profile description is missing, the username alone can be a valuable indicator. In one of their earlier papers, (Beskow and Carley, 2018) [29] focused on checking if a Twitter screen-name is random or not. They employed four features: character bi-grams, no. of lower case letters, no. of upper case letters and Shannon's string entropy. Although tested on a small number of accounts (100), their method shows promising results. However, since the accounts used for testing were hand-picked and were very few in number, and also taking into consideration the evolving nature of bots, it can be argued that, in the future, the randomness of a screen name might not be such a powerful feature as to be used alone in bot detection models.

4.2.2 Behavioral features

Behavioral features, which are content and language features based on twitting characteristics, have also been used in literature with great success[30], [22] [23]. Most notable features include: hashtags, mentions, length and entropy of tweets, shared links (URLs) or use of certain keywords. Additionally, prior research on temporal patterns suggests that they may uncover information about online campaigns and their evolution [31] [8]. As such, temporal features have also become frequently used in detection algorithms. Temporal features might include: average number of tweets produced over a period of time (day/week/the short period before an event), standard deviation of number of tweets over a time period and retweet rate (number of retweets out of all tweets).

Furthermore, language distribution statistics from a user's timeline have been used in the past in order to detect suspicious behavior. Certainly, it is more than normal for users to tweet in another language(s) than the one originally selected, considering the percentage of multilingual people across the globe. However, bots and trolls have been observed to be using an abnormal number of languages in what they tweet or retweet. This can be explained by the very nature of their mission, spreading misinformation. In order to reach as many users as possible, a multilingual approach might seem necessary. When looking for Russian trolls in the U.S. election of 2016, (Im et. al, 2019)[17] found that, when computing the intersection of all the languages used by the Russian accounts and control accounts, a surprising number of 82 languages resulted.

Finally, sentiment analysis has been used in some instances as a way to detect suspicious behavior. Research has shown how powerful a tool sentiment analysis can be, suggesting that sentiments extracted from online conversations were able to predict offline events, such as financial market fluctuation [32]. Moreover, attitude extracted from a piece of text is also known to affect information spreading [33].

4.2.3 Selected features

It should also be mentioned that network features have also been used in various models with success. Network features usually include three main types of networks: retweet networks, mention networks and hashtag co-occurrence networks. Those networks structures can give even more insight into the behavioral patterns of bots and the botnets they might be part of. However, given the time to collect each follower's timeline and the very large number of features we should extract for the analysis to be relevant, we decided to not use this information and focus on building a fast and easy to train classifier.

Few papers include information on the performance of each individual feature used in their bot detection model. Therefore, evaluating the most useful features requires testing. As such, we will initially start with using many features and, upon further analysis of the impact they have on detection (by using PCA to test the relevance and importance), gradually remove the redundant ones.

Table 1: Features used for the baseline model

Profile features

-screen name length	-username length	-username randomness (binary)
-account age	-default profile (binary)	-default picture (binary)
-profile description length	-no of followers	-no of followings
-ratio of followers-to-following	-statuses count	-background image(binary)

Behavioral features

-no of languages	-average length of tweets	-entropy of words in a tweet
-tweets count	-retweets count	-mentions counts
-retweets per tweet	-likes per tweet	-replies count
-average time between tweets	-average time between retweets	-URLs count
-average no of hashtags	-favourites count	-favourite-tweet ratio
-sentiment scores of aggregated tweets	-ratio of tweets containing emoticons	

5 Evaluation

It is import to acknowledge the limitations of the data used. The accounts were labelled by humans, so our results are hinging on correct annotations. It is entirely plausible that some well designed bots might have fooled even the most vigilant human annotator. Thus, when interpreting results, we must take this into consideration.

In order to evaluate our model, we will first use a 10-fold cross-validation method. The performance of the methodology will be evaluated in terms of recall and precision, with a focus on the precision value, since we especially want to avoid flagging a normal account as suspicious. We will informally define k-fold cross-validation, recall and precision as follows:

- **K-fold cross-validation**: the idea behind cross-validation is to create k non-overlapping folds (train/test chunks of data), so that each data point is used in the testing set one time, and is used to train the model k-1 times. The major benefit cross-validation brings is that it reduces the likelihood of choosing a biased or optimistic method, instead providing a good estimate of the model's performance on unseen data. The trade-off is that it's more computationally expensive, so it's less practical on very large sets of data. In the field of machine learning, values of k=5 or k=10 are very common, being the recommended default choices as they have been empirically shown to give an estimated error rate that does not suffer from very high variance of from very high bias. As such, we will use a 10-fold cross-validation approach.
- **Precision**: is a metric that indicates the fraction of retrieved accounts which are indeed relevant (bots). The higher the precision, the less human accounts we misclassify as being bots.
- **Recall**: is a metric that indicates the fraction of relevant accounts (bots) which are in our result set.

We will evaluate the performance of each model by taking into account the previously mentioned metrics, as well as provide details about the importance of each feature or set of features. As for the outcome, it is naive to assume that this could be an end-all solution. In the topic of bot detection, this model will only alleviate the issue by detecting a subset of all bots, with a focus on political bots.

6 Ethical implications and Risk management

Ethics-wise, the data was all collected following Twitter's Terms and Conditions and passed clearance from a University of Edinburgh ethics board (Social and Political Science). When the dataset will finally be published as open access on the UK Dataservice, it will also have been ethically verified by the UK Research Council. However, at least one ethical concern still remains: given the arm race between bots and misinformation, will our eventual success lead to counter measures by the bot creators? In the past years literature has shown that, indeed, bots have been continuously adapting to new detection algorithms, with varying degrees of success. As such, at the end of our project we will have to carefully consider whether providing too much information about our findings will potentially do more harm than good for the bot wars that social networks are currently fighting.

7 Research Plan, Milestones and Deliverables

- Literature review and further research: 10 days
- Clean and Process data: 5 days
- Exploratory data analysis: 5 days
- Build baseline model: 14 days
- Evaluate results, re-run analysis where necessary: 7 days
- Enrich methodology, improve baseline model: 20 days
- Write report: on-going
- Finalize report: 14 days

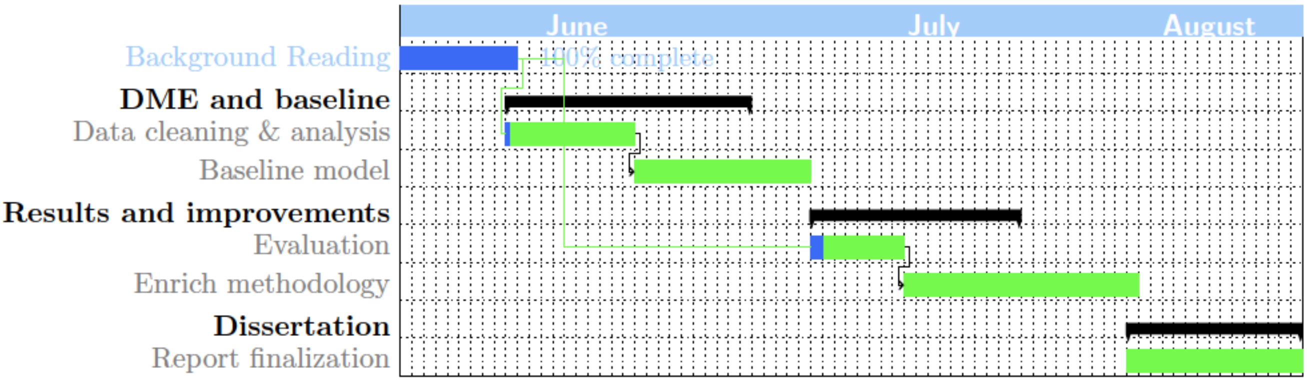

Figure 2: Gantt Chart of the activities defined for this project.

Milestone	**Week**	**Description**
M_1	2	Feasibility study completed
M_2	3	Data pre-processing and analysis completed
M_3	5	Baseline model implementation completed
M_4	6	Evaluation completed
M_5	9	Methodology enriched
M_4	11	Submission of dissertation

Table 2: Milestones defined in this project.

Deliverable	**Week**	**Description**
D_1	6	Baseline model for bot detection
D_2	7	Evaluation report
D_3	11	Dissertation

Table 3: List of deliverables defined in this project.

8 Extensions

Possible model extensions are listed below, however, these preferences may change over the course of the project:

- Check for stolen or shared profile image.
- Check for mismatch between twitter handles and screen name.
- Check for mismatch between profile picture gender and gender suggested by screen name.
- Check the performance of the model on another bot-related Twitter dataset.
- Analyze the media upload (in bytes) of bots vs humans (as suggested by [9]) and check if it would make for a relevant feature for the classifier.

References

[1] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *CoRR*, abs/1703.03107, 2017.

[2] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, 21, 11 2016.

[3] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from brexit and uselection. Working Paper 24631, National Bureau of Economic Research, May 2018.

[4] Philip Howard and Bence Kollanyi. Bots, strongerin, and brexit: Computational propaganda during the uk-eu referendum. *SSRN Electronic Journal*, 06 2016.

[5] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *CoRR*, abs/1707.00086, 2017.

[6] Lisa-Maria N. Neudert. Computational propaganda in germany: A cautionary tale. In *Oxford, UK: Project on Computational Propaganda*. Association for Computing Machinery, 2017.

[7] Norah Abokhodair, Daisy Yoo, and David W. McDonald. Dissecting a social botnet: Growth, content and influence in twitter. *CoRR*, abs/1604.03627, 2016.

[8] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Detection of promoted social media campaigns. In *Tenth International AAAI Conference on Web and Social Media*, page 563–566, 2016.

[9] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, and Jon Crowcroft. A large-scale behavioural analysis of bots and humans on twitter. *ACM Trans. Web*, 13(1), February 2019.

[10] Clare Llewellyn and Laura Cram. Distinguishing the wood from the trees: Contrasting collection methods to understand bias in a longitudinal brexit twitter dataset. In *Eleventh International AAAI Conference on Web and Social Media*. The AAAI Press, 5 2017.

[11] Clare Llewellyn, Laura Cram, Robin L. Hill, and Adrian Favero. For Whom the Bell Trolls: Shifting Troll Behaviour in the Twitter Brexit Debate. *Journal of Common Market Studies*, 57(5):1148–1164, September 2019.

[12] B. Viswanath, Ahmad Bashir, M. Crovella, S. Guha, Krishna P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. *Proceedings of the 23rd USENIX Security Symposium (USENIX Security)*, pages 223–238, 01 2014.

[13] Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Wang. Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73, 03 2014.

[14] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. 12 2016.

[15] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *CoRR*, abs/1802.04289, 2018.

[16] Chia-Mei Chen, D. J. Guan, and Qun-Kai Su. Feature set identification for detecting suspicious urls using bayesian classification in social networks. *Information Sciences*, 289:133–147, 12 2014.

[17] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. Still out there: Modeling and identifying russian troll accounts on twitter. *CoRR*, abs/1901.11162, 2019.

[18] Sangho Lee and Jong Kim. Early filtering of ephemeral malicious accounts on twitter. *Computer Communications*, 54, 08 2014.

[19] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 273–274, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[20] Alan Mislove, Sune Lehmann Jørgensen, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 554–557. AAAI Press, 2011.

[21] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, June 2016.

[22] Bo Wang, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Making the most of tweet-inherent features for social spam detection on twitter. volume 1395, 05 2015.

[23] Denis Stukal, Sergey Sanovich, Richard Bonneau, and Joshua Tucker. Detecting bots on russian political twitter. *Big Data*, 5:310–324, 12 2017.

[24] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. *CoRR*, abs/1811.03130, 2018.

[25] Kurt Thomas, Chris Grier, and Vern Paxson. Adapting social spam infrastructure for political censorship. pages 13–13, 04 2012.

[26] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *CoRR*, abs/1802.04291, 2018.

[27] Johan Fernquist, Lisa Kaati, and Ralph Schroeder. Political bots and the swedish general election. pages 124–129, 11 2018.

[28] David Beskow and Kathleen Carley. Its all in a name: Detecting and labeling bots by their name. volume 25, 12 2018.

[29] David Beskow and Kathleen Carley. *Using Random String Classification to Filter and Annotate Automated Accounts*, pages 367–376. 06 2018.

[30] Zhouhan Chen, Rima Tanash, Richard Stoll, and Devika Subramanian. Hunting malicious bots on twitter: An unsupervised approach. pages 501–510, 09 2017.

[31] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Identifying correlated bots in twitter. 11 2016.

[32] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[33] Emilio Ferrara and Zeyao Yang. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1, 06 2015.