Characterizing Social Spambots by their Human Traits

Salvatore Giorgi
University of Pennsylvania
sgiorgi@sas.upenn.edu

Lyle Ungar
University of Pennsylvania
ungar@cis.upenn.edu

H. Andrew Schwartz
Stony Brook University
has@cs.stonybrook.edu

Abstract

Social spambots, an emerging class of spammers attempting to emulate people, are difficult for both human annotators and classic bot detection techniques to reliably distinguish from genuine accounts. We examine this human emulation through studying the human characteristics (personality, gender, age, emotions) exhibited by social spambots' language, hypothesizing the values for these attributes will be unhuman-like (e.g. unusually high or low). We found our hypothesis mostly disconfirmed — *individually,* social bots exhibit very human-like attributes. However, a striking pattern emerged when consider the full distributions of these estimated human attributes: social bots were extremely similar and average in their expressed personality, demographics, and emotion (in contrast with traditional bots which we found to exhibit more variance and extreme values than genuine accounts). We thus consider how well social bots can be identified only using the 17 variables of these human attributes and ended up with a new state of the art in social spambot detection (e.g. $F1 = .946$). Further, simulating the situation of not knowing the bots *a priori*, we found that even an unsupervised clustering using the same 17 attributes could yield nearly as accurate of social bot identification ($F1 = 0.925$).

1 Introduction

A *social spambot* is "a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behavior" (Ferrara et al., 2016). Previous studies have shown that standard spambot detection algorithms as well as human annotators, while quite effective at detecting standard spam (e.g. sales advertisements), fail to accurately distinguish these human emulating *social spambots* from genuine Twitter accounts (Cresci et al., 2017).

In light of the goal of social spambots to emulate human behavior, we test the assumption that social spambots should not behave differently than human accounts and that a shared set of traits *should* be common among both groups. Thus, in this study we attempt to characterize and classify social spambots, in comparison to genuine Twitter accounts and traditional spambot accounts, by real human traits and states. We estimate demographics (age and gender), personality traits (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism), sentiment (positive and negative), and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) using a variety of pre-trained models. First and foremost, we describe the distributions of these 17 human attributes, and then we evaluate them as features in relatively simple spambot detection classifiers (random forest) and compare to more sophisticated state of the art detection methods.

Contributions Our contributions: (1) We characterize social spambots in terms of estimated human traits: demographics, personality, sentiment, and emotion. We show that social spambots express limited gender, age, and emotional variation while being much higher in positive sentiment and neurotic language than genuine users. (2) We show that traditional bots exhibit wider variance and more extreme estimated human traits. (3) We use these 17 human traits in a social spambot detection model, achieving state of the art classification results, in addition to outperforming human annotators.

2 Related Work

Bots and bot detection methods have received increased attention due to their role in spreading information, both real and fake, on social media platforms (Shao et al., 2018; Caldarelli et al., 2020). Particular focus has been given to their role in

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5148–5158

the 2016 U.S. Presidential election (Bessi and Ferrara, 2016; Badawy et al., 2018), but bots have been discovered discussing other political events such as the Syrian war (Abokhodair et al., 2015) and Brexit (Bastos and Mercea, 2019). Additionally, the spread of misinformation from bots has been linked to multiple public health issues such as vaccines (Broniatowski et al., 2018; Yuan et al., 2019), e-cigarettes (Allem et al., 2017), and marijuana (Allem et al., 2020). There is also growing evidence that bots are spreading information about COVID-19 (Himelein-Wachowiak et al., 2021; Ferrara, 2020; Al-Rawi and Shukla, 2020). The prevalence of bots in online discourse around elections and public policy has lead some researchers to begin outlining government policy for dealing with bots (Pedrazzi and Oehmer, 2020).

Bots have had such far reaching societal impact that research has grown significantly over the last decade (Cresci, 2020), with government agencies sponsoring competitions to identify the influence of bots (Subrahmanian et al., 2016). Although much bot research was done in the context of traditional (non-social or content generating) spambots, interest in social spambots has emerged more recently mostly focused on content, online behavior, or network attributes, rather than the human-likeness of the bots (Zhang et al., 2016). For example, Kudugunta and Ferrara (2018) applied deep neural architectures to the task of classifying social spambots from a single tweet. Using an adversarial learning approach Cresci et al. (2019b) generated evolved versions of current social spambots and, subsequently, attempted to classify them as bot or human. They again show that previous methods, which have been successful in identifying traditional bots, fail to detect the evolved social bots, including their own previous state-of-the-art method. Other studies have used anomaly detection (Miller et al., 2014) as well as sentiment based methods (Dickerson et al., 2014).

Our work is aligned with a growing set of methods to to embed language processing within the social and human contexts they are applied (Lynn et al., 2019). Most similar is the work on language generation or dialog agents (i.e. chatbots). While not directly related to spambot detection, such work has attempted to produce agents with empathy (Rashkin et al., 2019), trust (Novick et al., 2018) and emotion (Zhou and Wang, 2018; Huber et al., 2018) as well as general personalizations (Li et al., 2016; Zhang et al., 2018; Mazaré et al., 2018). As researchers build machines imbued with more sophisticated human attributes, we can expect similar machines to be used for spamming purposes.

3 Data Sets

We sought to use a variety of Twitter bot corpora in order to cover multiple contexts in which bots have been used. These contexts included social and non-social spambots (i.e., content generators and fake followers). In total, we use two social spambot data sets and four traditional bot data sets, each of which is briefly described below. While the focus of the current paper is social spambots, we investigate the human traits of traditional bots to show (1) that our methods generalize across different types of bots and (2) show that social spambots are indeed more sophisticated than traditional bots.

Spambot Data We use the two social spambot data sets derived from Cresci et al. (2017). The first data set, *SSB1* (Social SpamBots #1), is identical Cresci et al's test data: 464 social spambots, who advertised products on *Amazon.com*, plus 464 genuine accounts (718,975 total tweets). As a second data set, *SSB2*, we use an additional 2,913 genuine users from Cresci et al that were not part of *SSB1*, as well as 2,913 randomly selected social spambot accounts that Cresci found to be promoting a VIP version of the *Talnts* app. Because the number of tweets from the genuine accounts was much larger than that of the social spambots, we randomly sampled the genuine tweets, so that the tweet set was evenly split (2,621,684 total tweets). Appendix A contains a selection of social bot tweets to demonstrate how realistic they seem.

Traditional spambots We apply human trait estimates to four open source bot data sets, which consist of non-social bots (i.e., traditional / content generating bots and fake followers), in order to determine if these features generalize across data sets and bot types. All data sets are available through the Bot Repository [1]. We briefly summarize the data sets here (see source papers for more details) which contain both genuine and bot accounts. Yang et al. (2020): A data set of self-identified Twitter bots and verified accounts. Cresci et al. (2019a): Twitter accounts with suspicious financial tweets promoting low-value stocks. Cresci et al. (2015):

[1] https://botometer.osome.iu.edu/bot-repository/

Fake Twitter followers created to inflate the number of followers of popular accounts. Lee et al. (2011): A collection of content polluting Twitter accounts discovered via honeypot traps.

4 Estimating Human Traits

For each user in our data sets we estimate the following traits from their tweets: age, gender, personality, sentiment, and emotion.

Age / Gender. We applied a predictive lexicon to produce real valued age and gender estimates (Sap et al., 2014). This lexicon was built over Twitter, Facebook, and blog users with labeled age and gender, and produced prediction accuracies (Pearson r) of .831 (age) and .919 (gender).

Personality. We used a language based personality model to estimate the Big Five personality traits: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism (Park et al., 2015). This model was built over 1-3grams and a set of 2,000 LDA (Latent Dirichlet Allocation; Blei et al. 2003) topics and resulted in prediction accuracies (Pearson r) of .43 (openness), .37 (conscientiousness), .42 (extraversion), .35 (agreeableness) and .35 (neuroticism). Park et al. showed that these prediction accuracies are higher than correlations between self-report personality and personality ratings by friends.

Sentiment. We used the positive/negative sentiment categories from the NRC Word-Emotion Association Lexicon, a crowd sourced, word level lexicon (Emolex; Mohammad and Turney 2013).

Emotion. To estimate emotion we used the NRC Hashtag Emotion Lexicon (Mohammad and Kiritchenko, 2015), which has categories based on Plutchik's eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust (Plutchik, 1980). This is an automatically created, word-level lexicon based on tweets with emotion hashtags, i.e. *#anger*, *#anticipation*, etc.

5 Methods

Features For both the *SSB1* and *SSB2* data sets we extract all features needed to estimate the human traits. The age, gender, sentiment, and emotion lexica all use unigrams. For personality, we extract 1-3grams and a set of 2,000 LDA topics. These topics have been used across a number of studies on age, gender, and personality (Schwartz et al., 2013; Park et al., 2016) and were built over the MyPersonality data set (Kosinski et al., 2015).

	Technique	F1	Prec	Recall	Accuracy	MCC
Past Work	Humans	.570	.647	.509	.829	.470
	BotOrNot? (2016)	.761	.635	.950	.922	.738
	Ahmed et al. (2013)	.923	.913	.935	.923	.847
	Cresci et al. (2017)	.923	**1.000**	.858	.929	.867
This Work	Age & Gender	.578	.585	.582	.581	.167
	Personality	.899	.900	.899	.899	.800
	Sentiment	.833	.850	.835	.836	.684
	Emotions	.331	.248	.500	.500	.000
	All Human Traits	**.946***	.946	**.946***	**.946***	**.892**

Table 1: Classification results. First 4 lines presented in Cresci et al. (2017), bottom 5 lines are models presented in this work. * significantly different than Cresci et al. from a bootstrapped p-value < 0.01. MCC is the Matthews correlation coefficient.

5.1 Social Spambot Classification

Here the *SSB1* and *SSB2* data sets are used to train and test our model, respectively. In order to keep modeling simple, we built a random forest classifier, utilizing extremely randomized trees (Geurts et al., 2006), as implemented in by scikit-learn (Pedregosa et al., 2011). Modeling parameters are listed in Appendix D .

Comparisons We compare our model accuracies to human annotators and current state of the art methods, which are evaluated on our test set (*SSB2*) and reported in Cresci et al. (2017).

Human Annotators: a crowd sourcing task to test whether humans are able to distinguish social spambots from genuine accounts (Cresci et al., 2017).

BotOrNot: a supervised method trained to identify social spambots (Davis et al., 2016). The model uses over 1,000 different features including tweet content, sentiment and user meta data.

Ahmed et al.: an unsupervised graph clustering of features such as URLs, hashtags, mentions, and retweets (Ahmed and Abulaish, 2013).

Cresci et al.: uses DNA-inspired techniques to model behavior of Twitter users in an unsupervised fashion (Cresci et al., 2016).

5.2 Traditional Spambot Classification

In order to access the generalizability of our features, we evaluate a similar classification task using three open source bot data sets, which we briefly describe below (see the source papers for more details). In the previous task we had dedicated train and test data sets, in order to directly compare against the results in Cresci et al. (2017). Here, we do not have a consistent baseline therefore use a 10-fold cross validation setup, and use a model built on the top 1,000 most frequent unigrams as

a comparison. Again, we use an extremely randomized trees classifier; see Appendix E for full modeling and language processing details.

5.3 Unsupervised Classification

Finally, we apply an unsupervised clustering to our data in order to label each account as bot or human. We apply spectral clustering (Von Luxburg, 2007) to all 17 features (age, gender, personality, emotion, and sentiment) as well as the 5 personality dimensions and a baseline of 1,000 unigrams. The spectral clustering algorithm is implemented in scikit-learn with default settings. To simulate a more practical setting, where the human-to-bot ratio is not 50/50, we apply the clustering method to a 90/10 human-to-bot ratio. We fix the number of human accounts to the maximum number in both *SSB1* and *SSB2* (i.e., 2,913 and 464, respectively), and randomly sample bots so as to make a 90/10 human-to-bot ratio. For the 50/50 and 90/10 ratio we set the number of clusters to 2 and 10, respectively. We use 10 clusters on the unbalanced data since spectral clustering assumes equal cluster sizes. To chose the label for the bot cluster, we calculate the standard deviation of each human traits for each cluster, average the standard deviations within each cluster, and label the cluster with the minimum average standard deviation as "bots".

6 Results

Demographic, personality, sentiment, and emotion distributions are shown in Figures 1 and 2. The age estimates for the spambots (red) are tightly clustered (age mean = 28.9, standard deviation = 2.58), when compared to the genuine accounts (blue; mean = 23.8, standard deviation = 5.14). A similar pattern holds for gender – mean (standard deviation) gender is -.50 (.53) and .26 (1.17) for spambots and genuine accounts, respectively. The other traits show patterns similar to age and gender: social spambots have little variation in personality,

	Number of Humans/Bots	Feats.	F1	Prec.	Recall
Yang et al. (2020)	1971 / 670	unigrams	.978	.985	.971
		human traits	.968	.979	.957
Cresci et al. (2019a)	584 / 5,022	unigrams	.941	.921	.966
		human traits	.923	.928	.918
Cresci et al. (2015)	1,075 / 297	unigrams	.921	.969	.888
		human traits	.901	.948	.870
Lee et al. (2011)	17,720 / 14,632	unigrams	.851	.854	.849
		human traits	.862	.865	.862

Table 2: Traditional bot cross validation classification results for 1,000 unigrams and 17 human traits.

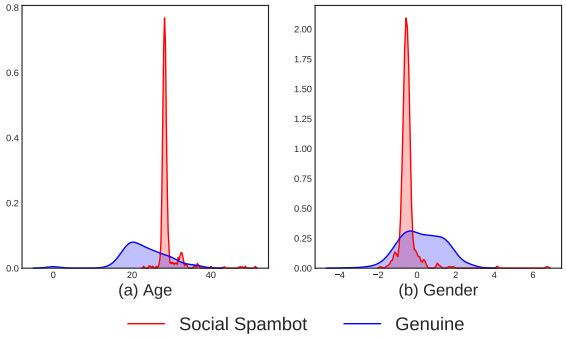

Figure 1: Age and gender distributions of genuine (blue) and social spambot (red) accounts in the *SSB2* data set.

sentiment, and emotion. Of note are neuroticism and positive sentiment distributions.

Classification results are presented in Table 1. The first four lines ("Past Work") contain results evaluated in Cresci et al. 2017 and are included here for context, while the current study's results are presented in the final five lines. Personality out performs all other single feature sets (age/gender, sentiment and emotion) across all of our evaluation metrics, while emotion performs the worst. Combining all features gives the best classification performance, significantly ($p < 0.01$) above the state of the art in Cresci et al. 2016.

Next, we evaluate estimated human traits for distinguish bots in a number of additional open source bot data sets. Since existing models were largely unavailable for these, we compare to a a much larger model built on the 1,000 most frequent unigrams. Cross validation classification results are in Table 2. The 17 human traits perform almost as well as the 1,000 unigram features, outperforming unigrams in the Lee et al. (2011) data set. The unigram models use (1) features which vary across each data set (i.e., the frequency of each unigram is calculated *within* each data set) and (2) a much larger number of features (1,000 vs. 17). This strongly suggests that the human traits are both generalizable across data sets and also bot types.

Finally, Table 3 shows the results of the unsupervised clustering task. Here, we see that even without providing the means for the models to learn the ideal ranges of attributes (i.e. from training data), the social bots primarily come out in one class with the genuine accounts in the other, even when we sample for a higher ratio of genuine-to-bot accounts (90/10).

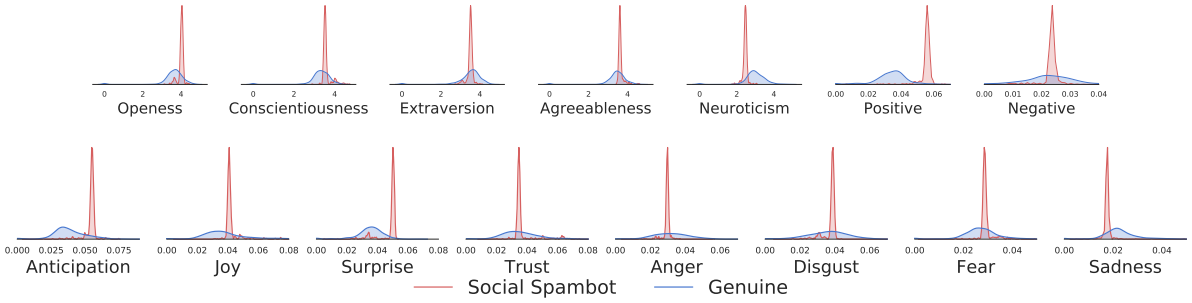

Figure 2: Personality and Emotion distributions of genuine (blue) and social spambot (red) accounts in the *SSB2* data set. Social spambots exhibited reasonable values but with little variance.

7 Ethics Statement

All bot data used in this study are publicly available. This study was reviewed by an academic institutional review board and found to be exempt, non-human subjects data. Still, ethical considerations should be raised when estimating demographics and personality from social media. These include privacy issues, biases in training data, the impact of misclassifications on downstream tasks, and outdated definitions of social constructs such as binary gender. While imperfect, we believe that these estimates are non-obtrusive and allow researchers to study average differences in estimated demographics as expressed in public language.

Further consideration was taken into account given the effectiveness of this approach for bot detection. Notably there is likely a continued arms race between bot creators and bot detectors: as natural language generation methods and bots become more advanced so too will detection methods. When published, it should be assumed that bot creators will try to reverse engineer the methods. Under open security principles, we hope the results of this study promote better detection, but caution that the effectiveness of the approach may ween over time.

		50/50 split			90/10 split		
		F1	Prec.	Recall	F1	Prec.	Recall
	MFC	.333	.248	.500	.474	.450	.500
SSB1	Unigrams	.334	.750	.500	.695	.665	.842
	Personality	**.832**	**.836**	**.832**	**.747**	**.712**	**.812**
	All Human Traits	.639	.803	.677	.472	.450	.498
SSB2	Unigrams	.336	.748	.502	.469	.450	.490
	Personality	.922	.931	.923	.949	.923	.980
	All Human Traits	**.925**	**.934**	**.926**	**.971**	**.995**	**.951**

Table 3: Unsupervised clustering classification results for both social spambot data sets using Most Frequent Class (MFC), personality, and all human traits.

8 Conclusion

In this study we showed that social spambots, while difficult for humans to detect, suffer from a lack of variation across all of our measures: age, gender, personality, sentiment, and emotion. Not only did social spambots seem to lack variation in most human traits, their mean values often aligned with the means of genuine accounts (with the exceptions of high mean positive sentiment and low neuroticism – both still with little variance).

The bots' lack of variation across any trait suggests they are "clone"-like. This could explain the inability of human annotators to properly identify them: *individual* social spambot accounts may look human (e.g., 29 years old and slightly male) but the range of humanness is limited. On the *population level*, most social spambots appear to be a clone of the same "human". This contrasts with traditional bots, whose traits were more varied and often with extreme, inhuman values (e.g., negative ages; see Appendix C), suggesting a "robot"-like interpretation of traditional bots.

The consistency of the estimated human traits was strong enough that we were able to build simple yet highly predictive classifiers using only the 17 estimates as features, outperforming human annotators and significantly more accurate than the state of the art in Cresci et al. (2017). On additional data sets spanning other types of bots (e.g., fake followers, spamming accounts, and self identifying bots), the human traits performed on par with much larger models (1,000 unigrams). Finally, simulating where bot training data is non-existant, we showed a simple unsupervised clustering based on the 17 attributes identified social bots nearly as accurately, suggesting human trait based approaches may be able to identify new bots in the wild.

References

Norah Abokhodair, Daisy Yoo, and David W McDonald. 2015. Dissecting a social botnet: Growth, content and influence in twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 839–851.

Faraz Ahmed and Muhammad Abulaish. 2013. A generic statistical approach for spam detection in online social networks. *Computer Communications*, 36(10-11):1120–1129.

Ahmed Al-Rawi and Vishal Shukla. 2020. Bots as active news promoters: A digital analysis of covid-19 tweets. *Information*, 11(10):461.

Jon-Patrick Allem, Patricia Escobedo, and Likhit Dharmapuri. 2020. Cannabis surveillance with twitter data: emerging topics and social bots. *American journal of public health*, 110(3):357–362.

Jon-Patrick Allem, Emilio Ferrara, Sree Priyanka Uppu, Tess Boley Cruz, and Jennifer B Unger. 2017. E-cigarette surveillance with social media data: social bots, emerging topics, and trends. *JMIR public health and surveillance*, 3(4):e98.

Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 258–265. IEEE.

Marco T Bastos and Dan Mercea. 2019. The brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1):38–54.

Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11-7).

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384.

Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. 2020. The role of bot squads in the political propaganda on twitter. *Communications Physics*, 3(1):1–15.

Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2016. Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. 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*, pages 963–972. International World Wide Web Conferences Steering Committee.

Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi. 2019a. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, 13(2):1–27.

Stefano Cresci, Marinella Petrocchi, Angelo Spognardi, and Stefano Tognazzi. 2019b. Better safe than sorry: An adversarial approach to improve social bot detection. *arXiv preprint arXiv:1904.05132*.

Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee.

John P Dickerson, Vadim Kagan, and VS Subrahmanian. 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 620–627. IEEE.

Emilio Ferrara. 2020. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*.

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

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63(1):3–42.

McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, and Brenda Curtis. 2021. Bots and misinformation spread on social media: Implications for covid-19. *J Med Internet Res*, 23(5):e26933.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 277. ACM.

Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.

Sneha Kudugunta and Emilio Ferrara. 2018. Deep neural networks for bot detection. *Information Sciences*, 467:312–322.

Kyumin Lee, Brian Eoff, and James Caverlee. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H. Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28, Minneapolis, Minnesota. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

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

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

David Novick, Mahdokht Afravi, Adriana Camacho, Laura J Hinojos, and Aaron E Rodriguez. 2018. Inducing rapport-building behaviors in interaction with an embodied conversational agent. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 345–346. ACM.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Gregory Park, David Bryce Yaden, H Andrew Schwartz, Margaret L Kern, Johannes C Eichstaedt, Michael Kosinski, David Stillwell, Lyle H Ungar, and Martin EP Seligman. 2016. Women are warmer but no less assertive than men: Gender and language on facebook. *PloS one*, 11(5):e0155885.

Stefano Pedrazzi and Franziska Oehmer. 2020. Communication rights for social bots?: Options for the governance of automated computer-generated online identities. *Journal of Information Policy*, 10:549–581.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Maarten Sap, Greg Park, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, Lyle H Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Venkatramanan S Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. The darpa twitter bot challenge. *Computer*, 49(6):38–46.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293.

Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1096–1103.

Xiaoyi Yuan, Ross J Schuchard, and Andrew T Crooks. 2019. Examining emergent communities and social bots within the polarized online vaccination debate in twitter. *Social Media+ Society*, 5(3):2056305119865465.

Jinxue Zhang, Rui Zhang, Yanchao Zhang, and Guanhua Yan. 2016. The rise of social botnets: Attacks and countermeasures. *IEEE Transactions on Dependable and Secure Computing*, 15(6):1068–1082.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

A Spambot Sample Tweets

We include a set of random tweets from each social spambot data set, in order to show that these accounts are not immediately identifiable as bots.

	Tweet
SSB1	• To your favorite song, anything you want to do.. • Best week for a long time • Seriously 90210 ended on the best note. I was in waaay too many tears. What do I do with my life now?
SSB2	• I don't read other science fiction. I don't read any at all. - Jack Vance • Better to be a geek than an idiot. • I like my money right where I can see it – hanging in my closet. - Carrie Bradshaw

Table 4: Sample of random social spambot tweets, presented here to highlight the fact that the tweet authors are not immediately identifiable as non-human.

B Social Spambots: Training Data

Here we plot the distributions of our 17 features for the social spambots and genuine users in our training data (see main paper for test data). Demographic data (age and gender) are plotted in Figure 3. We see a similar pattern to the test data, though less pronounced: social spambot distributions are considerably more peaked than the genuine users. Also, note that genuine users have a bimodal gender distribution, corresponding to the female/male splot, whereas the spambot distribution is unimodal. Figure 4 shows the distributions for personality, sentiment and emotion. Again, we see patterns similar to the test data, though less pronounced. The train spambots are also more positive than the genuine users, though the emotional stability split, seen in the test data, is not as pronounced.

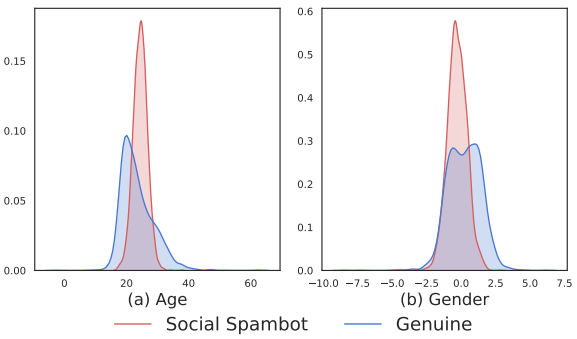

Figure 3: Demographic (age and gender) distributions of genuine and social spambot accounts in the training data.

C Human Traits: Traditional vs Social Spambots

We estimates the human traits of a set of 1,000 traditional spambots from Cresci et al. (2017), sampled from a larger set of data released in Yang et al. (2013). In Figure 5 we compare age and gender distributions to those of genuine Twitter users and social spambots. Here we see distinct distributions for each type of account, with traditional spambots having a much wider distribution. In the case of age, we see that traditional bots have non-human predictions (i.e., negative numbers), whereas the predicted age of social spambots is within a reasonable human range. The gender distribution of the traditional spambots do not show a female/male split (i.e., no distinct bimodal distribution). Thus, we have two distinct patterns: (1) traditional bots exhibit a wide range of characteristics, though often non-human and uninterpretable; and, (2) social bots exhibit a very narrow band of characteristics, though this band is within a normal human range, or at least within the range of predicted values for genuine accounts.

In Figure 6 we present the distributions for personality, sentiment, and emotion for all three classes of Twitter users: genuine accounts, traditional spambots and social spambots. First, we see three distinct distributions for each class of Twitter users across all characteristics. Similar to the age and gender plots, when comparing emotions to genuine users, the traditional bots often have a larger spread and are multi-modal. For personality and sentiment, we see a different pattern. Here the traditional bots are more spread than the social bots but still fairly limited when compared to genuine users. While the social spambots have a consistently different pattern when compared to genuine users, we see slightly more variation when comparing traditional bots to genuine users. That said, when comparing across all traits we see three fairly distinct classes of Twitter users.

D Experimental Setup for Classifying Social Bots

We used the following settings in our extremely randomized trees classifier: *bootstrap* is False, *class_weight* is None, *criterion* is gini, *max_depth* is None, *max_features* is sqrt, *max_leaf_nodes* is None, *min_impurity_split* is 1^{-7}, *min_samples_leaf* is 100, *min_samples_split* is 2, *min_weight_fraction_leaf* is 0, *n_estimators* is 1000,

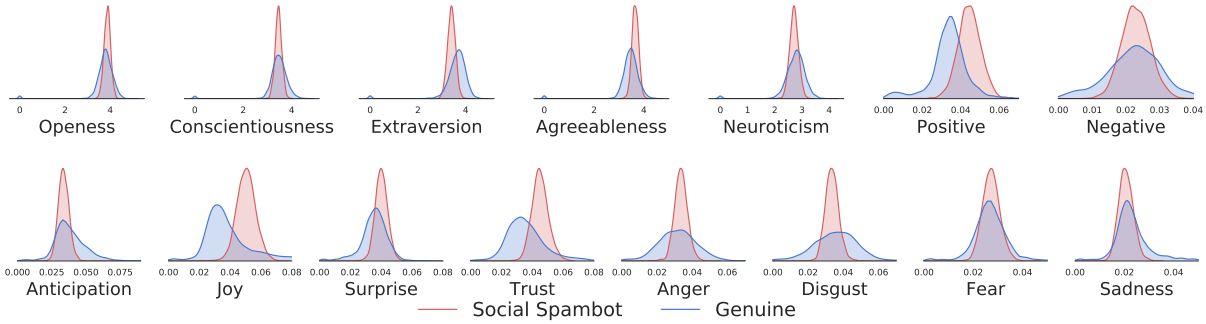

Figure 4: Personality, Sentiment and Emotion distributions of genuine accounts and social spambots in the training data.

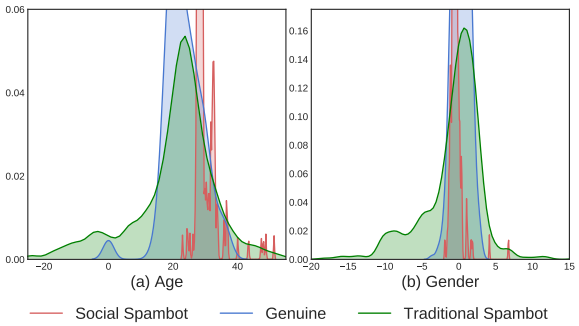

Figure 5: Demographic distributions of traditional spambots, genuine accounts and social spambots. Zoomed to highlight traditional spambots, see Figure 1 in the main paper for genuine accounts and social spambots.

n_jobs is 5, *oob_score* is False, *random_state* is None, and *warm_start* is False.

E Experimental Setup for Classifying Traditional Bots

We use our 17 human trait features in a classification task where we attempt to distinguish traditional bots from genuine human accounts. To do this we use four open source data sets: Yang et al. (2020), Cresci et al. (2019a), Cresci et al. (2015), and Lee et al. (2011). These data sets are all available at the Bot Repository[2], which is a centralized repository for open source bot data sets. Due to privacy reasons, the majority of the data sets on this repo do not contain tweet level data, which we need to estimate the human traits. Therefore, we limit our analysis to three data sets which have tweets available on the Bot Repository: Cresci et al. (2019a), Cresci et al. (2015), and Lee et al. (2011). For the fourth data set in our task, Yang et al. (2020), we use the Twitter API[3] to download available tweet histories from the *botwiki* and *verified* described in the paper (see Yang et al. (2020) for more details).

For each of these data sets we first apply an English filter, since our human trait estimators are trained on English social media data. We use use the langid Python package (Lui and Baldwin, 2012) to restrict to English only tweets. We then extract 1-3grams and a set of 2,000 LDA topics for each account, which are then used to estimate the human traits (age, gender, personality, emotion, and sentiment). See main paper for details on feature extraction. Since we are not using the original tweet set as reported in the above papers (due to English filtering or re-download the tweets from the Twitter API), we build a comparison model using the 1,000 most frequent unigrams. Finally, we restrict our analysis to accounts which have posted at least 500 words across their English tweet histories. To classify each account as human or bot we run 10-fold cross validation using stratified folds. We use an extremely randomized trees classifier as implement in scikit-learn using the following parameters using the settings listed in the social bot detection modeling. Two models are build, one for each of the following feature sets: (1) 1,000 most frequent 1-3grams and (2) 17 human traits.

Additional Acknowledgements

We wish to thank the anonymous reviewers as well as colleagues Joao Sedoc, Johannes Eichstaedt, and David Yaden for their valuable feedback on this work. Dr. Schwartz's effort was funded in part by the Defense Advanced Research Projects Agency (DARPA), via grant #W911NF-20-1-0306 to Stony Brook University. The conclusions and opinions

[2] https://botometer.osome.iu.edu/bot-repository/datasets.html

[3] https://github.com/dlatk/TwitterMySQL

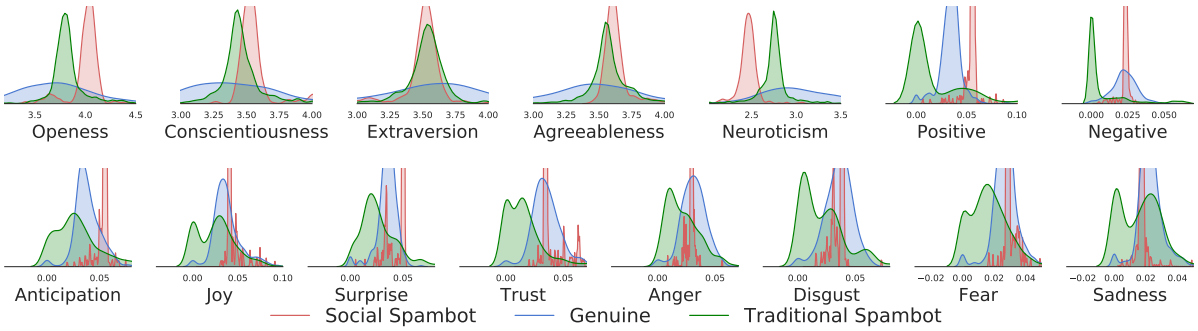

Figure 6: Personality, Sentiment and Emotion distributions of genuine accounts, traditional spambots and social spambot.

expressed are attributable only to the authors and should not be construed as those of DARPA or the U.S. Department of Defense.