Journal of Computational Science 16 (2016) 1–7

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Sifting robotic from organic text: A natural language approach for detecting automation on Twitter

Eric M. Clark[a,b,c,d,e,*], Jake Ryland Williams[a,b,c,d], Chris A. Jones[e,f,g], Richard A. Galbraith[h,i], Christopher M. Danforth[a,b,c,d], Peter Sheridan Dodds[a,b,c,d]

[a] Department of Mathematics & Statistics, University of Vermont, Burlington, VT 05401, United States
[b] Vermont Complex Systems Center, University of Vermont, Burlington, VT 05401, United States
[c] Vermont Advanced Computing Core, University of Vermont, Burlington, VT 05401, United States
[d] Computational Story Lab, University of Vermont, Burlington, VT 05401, United States
[e] Department of Surgery, University of Vermont, Burlington, VT 05401, United States
[f] Global Health Economics Unit of the Vermont Center for Clinical and Translational Science, University of Vermont, Burlington, VT, 05401, United States
[g] Vermont Center for Behavior and Health, University of Vermont, Burlington, VT 05401, United States
[h] Department of Medicine, University of Vermont, Burlington, VT 05401, United States
[i] Vermont Center for Clinical and Translational Science, University of Vermont, Burlington, VT 05401, United States

ARTICLE INFO

Article history:
Received 8 June 2015
Received in revised form 11 October 2015
Accepted 10 November 2015
Available online 19 November 2015

ABSTRACT

Twitter, a popular social media outlet, has evolved into a vast source of linguistic data, rich with opinion, sentiment, and discussion. Due to the increasing popularity of Twitter, its perceived potential for exerting social influence has led to the rise of a diverse community of automatons, commonly referred to as bots. These inorganic and semi-organic Twitter entities can range from the benevolent (e.g., weather-update bots, help-wanted-alert bots) to the malevolent (e.g., spamming messages, advertisements, or radical opinions). Existing detection algorithms typically leverage metadata (time between tweets, number of followers, etc.) to identify robotic accounts. Here, we present a powerful classification scheme that exclusively uses the natural language text from organic users to provide a criterion for identifying accounts posting automated messages. Since the classifier operates on text alone, it is flexible and may be applied to any textual data beyond the Twittersphere.

1. Introduction

Twitter has become a mainstream social outlet for the discussion of a myriad of topics through microblogging interactions. Members chiefly communicate via short text-based public messages restricted to 140 characters, called tweets. As Twitter has evolved from a simple microblogging social media interface into a mainstream source of communication for the discussion of current events, politics, consumer goods/services, it has become increasingly enticing for parties to gameify the system by creating automated software to send messages to organic (human) accounts as a means for personal gain and for influence manipulation [1,2]. The results of sentiment and topical analyses can be skewed by robotic accounts that dilute legitimate public opinion by algorithmically generating vast amounts of inorganic content. Nevertheless, data from Twitter is becoming a source of interest in public health and economic research in monitoring the spread of disease [4,5] and gaining insight into public health trends [6].

In related work [7–10], researchers have built classification algorithms using metadata idiosyncratic to Twitter, including the number of followers, posting frequency, account age, number of user mentions/replies, username length, and number of retweets. However, relying on metadata can be problematic: sophisticated spam algorithms now emulate the daily cycle of human activity and author borrowed content to appear human [7]. Another problematic spam tactic is the renting of accounts of legitimate users (called sponsored accounts), to introduce short bursts of spam and hide under the user's organic metadata to mask the attack [11].

A content based classifier proposed by Chu et al. [19] measures the entropy between Twitter time intervals along with user metadata to classify Twitter accounts, and requires a comparable number of tweets (≥60) for adequate classification accuracy as our proposed method. SentiBot, another content based

* Corresponding author at: University of Vermont, Burlington, VT 05401, United States
E-mail address: eclark@uvm.edu (E.M. Clark).

http://dx.doi.org/10.1016/j.jocs.2015.11.002
1877-7503/

classifier [20], utilizes latent Dirichlet allocation (LDA) for topical categorization combined with sentiment analysis techniques to classify individuals as either bots or humans. We note that as these automated entities evolve their strategies, combinations of our proposed methods and studies previously mentioned may be required to achieve reasonable standards for classification accuracy. Our method classifies accounts solely based upon their linguistic attributes and hence can easily be integrated into these other proposed strategies.

We introduce a classification algorithm that operates using three linguistic attributes of a user's text. The algorithm analyzes:

1 the average URL count per tweet,
2 the average pairwise lexical dissimilarity between a user's tweets,
and
3 the word introduction rate decay parameter of the user for various proportions of time-ordered tweets.

We provide detailed descriptions of each attribute in the next section. We then test and validate our algorithm on 1000 accounts which were hand coded as automated or human.

We find that for organic users, these three attributes are densely clustered, but can vary greatly for automatons. We compute the average and standard deviation of each of these dimensions for various numbers of tweets from the human coded organic users in the dataset. We classify accounts by their distance from the averages from each of these attributes. The accuracy of the classifier increases with the number of tweets collected per user. Since this algorithm operates independently from user metadata, robotic accounts do not have the ability to adaptively conceal their identities by manipulating their user attributes algorithmically. Also, since the classifier is built from time ordered tweets, it can determine if a once legitimate user begins demonstrating dubious behavior and spam tactics. This allows for social media data-miners to dampen a noisy dataset by weeding out suspicious accounts and focus on purely organic tweets.

2. Data handling

2.1. Data-collection

We filtered a 1% sample of Twitter's streaming API (the spritzer feed) for tweets containing geo-spatial metadata spanning the months of April through July in 2014. Since roughly 1% of tweets provided GPS located spatial coordinates, our sample represents nearly all of the tweets from users who enable geotagging. This allows for much more complete coverage of each user's account. From this sample, we collected all of the geo-tweets from the most active 1000 users for classification as human or robot and call this the Geo-Tweet dataset.

2.2. Social HoneyPots

To place our classifier in the context of recent work, we applied our algorithm to another set of accounts collected from the Social HoneyPot Experiment [12]. This work exacted a more elaborate approach to find automated accounts on Twitter by creating a network of fake accounts (called Devils [13]) that would tweet about trending topics amongst themselves in order to tempt robotic interactions. The experiment was analyzed and compiled into a dataset containing the tweets of "legitimate users" and those classified as "content polluters". We note that the users in this dataset were not hand coded. Accounts that followed the Devil HoneyPot accounts were deemed robots. Their organic users were compiled from a random sample of Twitter, and were only deemed organic because these accounts were not suspended by Twitter at the time. Hence the full HoneyPot dataset can only serve as an estimate of the capability of this classification scheme.

2.3. Human classification of Geo-Tweets

Each of the 1000 users were hand classified separately by two evaluators. All collected tweets from each user were reviewed until the evaluator noticed the presence of automation. If no subsample of tweets appeared to be algorithmically generated, the user was classified as human. The results were merged, and conflicting entries were resolved to produce a final list of user ids and codings. See Fig. 1 for histograms and violin plots summarizing the distributions of each user class. We note that any form of perceived automation was sufficient to deem the account as automated. See Supplementary Material for samples of each of these types of tweets from each user class and a more thorough description of the annotation process.

2.4. Types of users

We consider organic content, i.e. from human accounts, as those that have not tweeted in an algorithmic fashion. We focused on three distinct classes of automated tweeting:

Robots: Tweets from these accounts draw on a strictly limited vocabulary. The messages follow a very structured pattern, many of which are in the form of automated updates. Examples include Weather Condition Update Accounts, Police Scanner Update Accounts, Help Wanted Update Accounts, etc.

Cyborgs: The most covert of the three, these automatons exhibit human-like behavior and messages through loosely structured, generic, automated messages and from borrowed content copied from other sources. Since many malicious cyborgs on Twitter try to market an idea or product, a high proportion of their tweets contain URLs, analogous to spam campaigns studied on Facebook [14]. Messages range from the backdoor advertising of goods and services [15] to those trying to influence social opinion or even censor political conversations [16]. These accounts act like puppets from a central algorithmic puppeteer to push their product on organic users while trying to appear like an organic user [17]. Since these accounts tend to borrow content, they have a much larger vocabulary in comparison to ordinary robots. Due to Twitter's 140 character-per-tweet restriction, some of the borrowed content being posted must be truncated. A notable attribute of many cyborgs is the presence of incomplete messages followed by an ellipsis and a URL. Included in this category are 'malicious promoter' accounts [12] that are radically promoting a business or an idea systematically.

Human Spammers: These are legitimate accounts that abuse an algorithm to post a burst of almost indistinguishable tweets that may differ by a character in order to fool Twitter's spam detection protocols. These messages are directed at a particular user, commonly for a follow request to attempt to increase their social reach and influence.

Although we restrict our focus to the aforementioned classes, we did notice the presence of other subclasses, which we have named "listers", and "quoters", that have both organic and automaton features. Listers are accounts that send their messages to large groups of individuals at once. Quoters are dedicated accounts that are referencing distant passages from literature or song lyrics. Most of the tweets from these accounts are all encased in quotations. These accounts also separately tweet organic content. We classified these accounts as human because there was not sufficient evidence suggesting these behaviors were indeed automated.

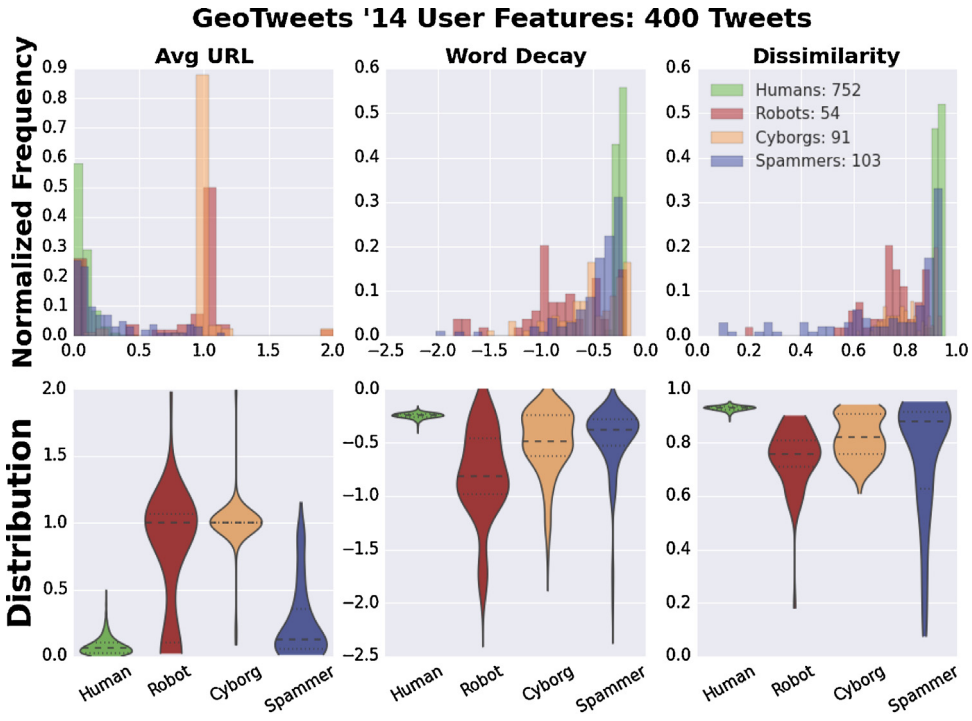

Fig. 1. The feature distribution of the 1000 hand coded users are summarized with histograms and violin plots. These show the wide variation in automated features versus Organics. Violin plots show the kernel density estimation of each distribution. Using the Organic features, automated entities are identified by exclusion.

3. Methods

3.1. Classification algorithm

The classifier, $\mathcal{C}$, takes ordinal samples of tweets from each user, μ, of varying number, s, to determine if the user is a human posting strictly organic content or is algorithmically automating tweets:

$$\mathcal{C} : \mu^s \to \{0, 1\} = \{\text{Organic, Automaton}\}.$$

Although we have classified each automaton into three distinct classes, the classifier is built more simply to detect and separate organic content from automated. To classify the tweets from a user, we measure three distinct linguistic attributes:

1 Average pairwise tweet dissimilarity,
2 Word introduction rate decay parameter,
3 Average number of URLs (hyperlinks) per tweet.

3.2. Average pairwise tweet dissimilarity

Many algorithmically generated tweets contain similar structures with minor character replacements and long chains of common substrings. Purely organic accounts have tweets that are very dissimilar on average. The length of a tweet, t, is defined as the number of characters in the tweet and is denoted $|t|$. Each tweet is cleaned by truncating multiple whitespace characters and the metric is performed case insensitively. A sample of s tweets from a particular user is denoted T^s_μ. Given a pair of tweets from a particular user, $t_i, t_j \in T^s_\mu$, the pairwise tweet dissimilarity, $D(t_i, t_j)$, is given by subtracting the length of the longest common subsequence of both tweets, $|LCS(t_i, t_j)|$ and then weighting by the sum of the lengths of both tweets:

$$D(t_i, t_j) = \frac{|t_i| + |t_j| - 2 \cdot |LCS(t_i, t_j)|}{|t_i| + |t_j|}.$$

The average tweet dissimilarity of user μ for sample size of s tweets is calculated as:

$$\mu^s_{lcs} = \frac{1}{(s-1)!} \cdot \sum_{t_i, t_j \in T^s_\mu} D(t_i, t_j).$$

For example, given the two tweets: (t_1, t_2) = (I love Twitter, I love to spam). Then $|t_1| = |t_2| = 14$, $LCS(t_1, t_2) = |\text{I love } t| = 8$ (including whitespaces) and we calculate the pairwise tweet dissimilarity as:

$$D(t_1, t_2) = \frac{14 + 14 - 2 \cdot 8}{14 + 14} = \frac{12}{28} = \frac{3}{7}.$$

3.3. Word introduction decay rate

Since social robots *automate* messages, they have a limited and crystalline vocabulary in comparison to organic accounts. Even cyborgs that mask their automations with borrowed content cannot fully mimic the rate at which organic users introduce unique words into their text over time. The word introduction rate is a measure of the number of unique word types introduced over time from a given sample of text [18]. The rate at which unique words are introduced naturally decays over time, and is observably different between automated and organic text. By testing many random word shufflings of a text, we define $\bar{m}_n$ as the average number of words between the nth and $n+1$st initial unique word type appearances. From [18], the word introduction decay rate, $\alpha(n)$, is given as

$$\alpha(n) = 1/\bar{m}_n \propto n^{-\gamma} \quad \text{for} \quad \gamma > 0.$$

For each user, the scaling exponent of the word introduction decay rate, α, is approximated by performing standard linear regression on the last third of the log-transformed tail of the average gap size distribution as a function of word introduction number, n [18]. In Fig. 2, the log transformed rank-unique word gap distribution is

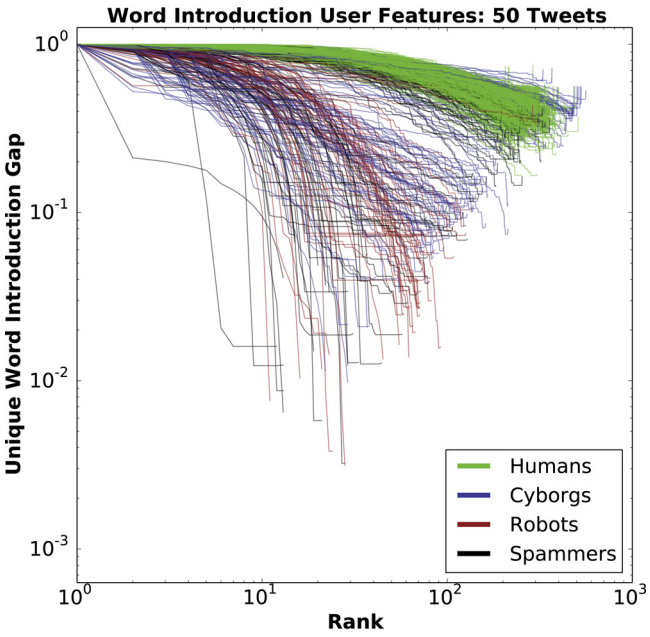

Fig. 2. The rank-unique word gap distribution is plotted on a logscale for each user class.

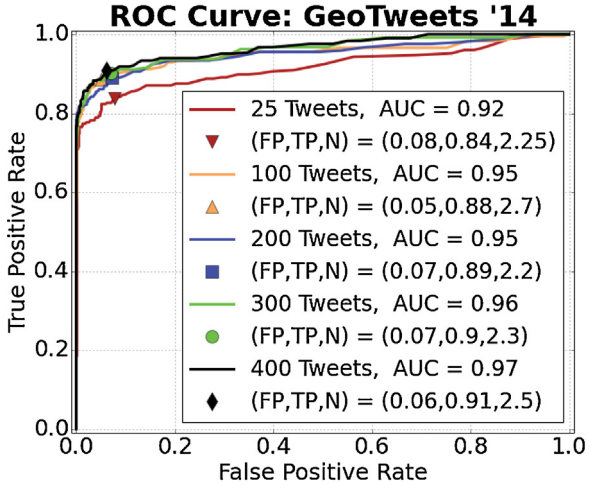

Fig. 3. The receiver operator characteristic curve from the 10-fold Cross Validation Experiment performed on the Geo Tweets collected from April through July 2014. The True Positive (TP), False Positive (FP), and thresholds, *N*, are averaged across the 10 trials. The accuracies are approximated by the AUCs, which we compute using the trapezoid rule. The points depict the best experimental model thresholding window (*N*).

given for each individual in the data set. Here the human population (green) is distinctly distributed in comparison to the automatons.

3.4. Average URLs per tweet

Hyperlinks (URLs) help automatons spread spam and malware [11,21,22]. A high fraction of tweets from spammers tend to contain some type of URL in comparison to organic individuals, making the average URLs per tweet a valuable attribute for bot classification algorithms [9,23,24]. For each user, the average URL rate is measured by the total number of occurrences of the substring 'http:' within tweets, and then divided by the total number of tweets authored by the user in the sample of size s:

$$\mu^{s}_{url} = \frac{\#\text{Occurrences of 'http:'}}{\#\text{Sampled Tweets}}.$$

3.5. Cross Validation experiment

We perform a standard 10-fold Cross Validation procedure on the 2014 Geo-Tweet data set to measure the accuracy of using each linguistic feature for classifying Organic accounts. We divided individuals into 10 equally sized groups. Then 10 trials are performed where 9 of the 10 groups are used to train the algorithm to classify the final group.

During the calibration phase, we measure each of the three features for every human coded account in the training set. We sequentially collect tweets from each user from a random starting position in time. We record the arithmetic mean and standard deviation of the Organic attributes to classify the remaining group. The classifier distinguishes human from automaton by using a varying threshold, n, from the average attribute value computed from the training set. For each attribute, we classify each user as an automaton if their feature falls further than n standard deviations away from the organic mean, for varying n.

For each trial, the False Positives and True Positives for a varying window size, n, are recorded. To compare to other bot-detection strategies, we rate True Positives as the success at which the classifier identifies automatons by exclusion, and False Positives as humans that are incorrectly classified as automatons. The results of the trials for varying tweet sizes are averaged and visualized with a Receiver Operator Characteristic curve (ROC) (see Fig. 3). The accuracy of each experiment is measured as the area under the ROC, or AUC. To benchmark the classifier, a 10-fold Cross Validation was also performed on the HoneyPot tweet-set which we describe in the following section.

4. Results and discussion

4.1. Geo-Tweet Classification Validation

The ROC curves for the Geo-Tweet 10 fold Cross Validation Experiment for varying tweet bins in Fig. 3 show that the accuracy increases as a function of number of tweets.

Although the accuracy of the classifier increases with the number of collected tweets, we see in Fig. 4 that within 50 tweets the accuracy of the average of 10 random trials is only slightly higher than a 500 tweet user sample. While this is very beneficial to our task (isolating humans), we note that larger samples see greater returns when one instead wants to isolate spammers, that tweet random bursts of automation.

4.2. HoneyPot external validation

The classifier was tested on the Social Honeypot Twitter-bot dataset provided by Lee et al. [12]. Results are visualized with a ROC curve in Fig. 5. The averaged optimal threshold for the full English user dataset (blue curve) had a high true positive rate (correctly classified automatons: 86%), but also had a large false positive rate (misclassified humans: 22%).

The Honeypot Dataset relied on Twitter's spam detection protocols to label their randomly collected "legitimate users". Some forms of automation (weather-bots, help-wanted bots) are permitted by Twitter. Other cyborgs that are posting borrowed organic content can fool Twitter's automation criterion. This ill formation of the training set greatly reduces the ability of the classifier to distinguish humans from automatons, since the classifier gets the wrong information about what constitutes a human. To see this, a random sample of 1000 English Honeypot users was hand-coded to mirror the previous experiment. On this smaller sample (black

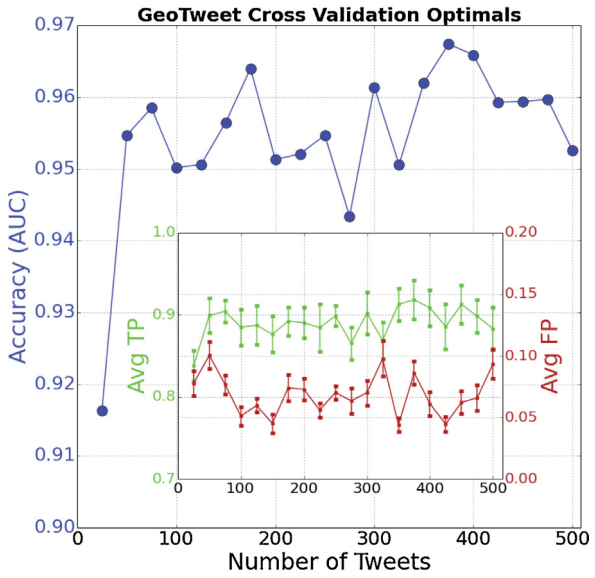

Fig. 4. Accuracy, computed as the AUC is plotted as a function of number of tweets, ranging from 25 to 500. The average True Positive and False Positive Rates over 10 trials is given on twin axes with error bars drawn using the standard error.

curve in Fig. 4), the averaged optimal threshold accuracy increased to 96%.

4.3. Calibrated classifier performance

We created the thresholding window of final calibrated classifier using the results from the calibration experiment. We average the optimal parameters from the 10-fold cross validation on the Geo-Tweet dataset from each of the 10 calibration trials for tweet bins ranging from 25 to 500 in increments of 25 tweets. We also average and record the optimal parameter windows, n_{opt} and their standard deviations, σ_{opt}. The standard deviations serve as a tuning parameter to increase the sensitivity of the classifier, by increasing the feature cutoff window (n). The results from applying the calibrated classifier to the full set of 1000 users, using 400 tweet bags is given in Fig. 6. The feature cutoff window (black lines) estimates if the user's content is organic or automated. Human feature sets

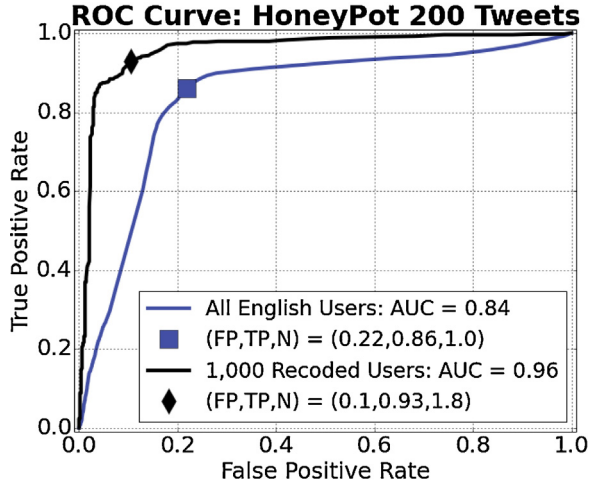

Fig. 5. HoneyPot Data Set, 10 fold Cross Validation Performance for users with 200 tweets. The black curve represents the 1000 hand coded HoneyPot users, while the blue curve is the entire English Honeypot dataset. The accuracy increases from 84% to 96%. (For interpretation of reference to color in this figure legend, the reader is referred to the web version of this article.)

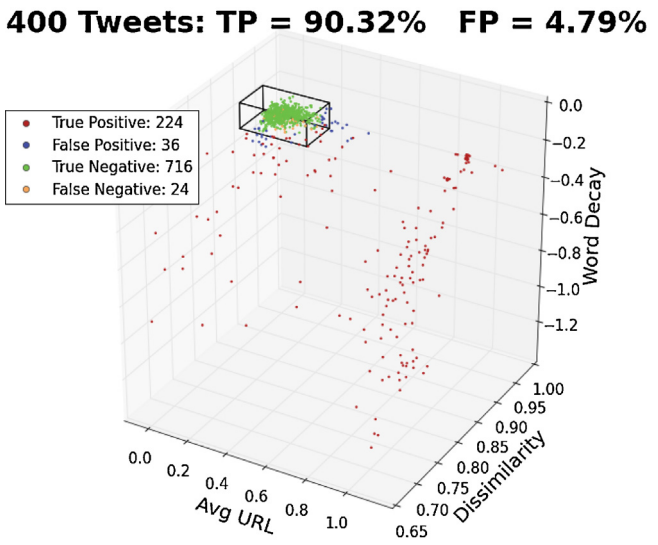

Fig. 6. Calibrated Classifier Performance on 1000 User Geo-Tweet Dataset. Correctly classified humans (True Negative), are coded in Green, while correctly identified automatons (True Positives) are coded in red. The 400 tweet average optimal thresholds from the cross validation experiment designate the thresholding for each feature. The black lines demonstrates each feature cutoff. (For interpretation of reference to color in this figure legend, the reader is referred to the web version of this article.)

(True Negatives: 716) are densely distributed with a 4.79% False Positive Rate (i.e., humans classified as robots). The classifier accurately classified 90.32% of the automated accounts and 95.21% of the Organic accounts. See Fig. S1 for a cross sectional comparison of each feature set. We note that future work may apply different methods in statistical classification to optimize these feature sets, and that using these simple cutoffs already leads to a high level of accuracy.

5. Conclusion

Using a flexible and transparent classification scheme, we have demonstrated the potential of using linguistic features as a means of classifying automated activity on Twitter. Since these features do not use the metadata provided by Twitter, our classification scheme may be applicable outside of the Twittersphere. Future work can extend this analysis multilingually and incorporate additional feature sets with an analogous classification scheme. URL content can also be more deeply analyzed to identify organic versus SPAM related hyperlinks.

We note the potential for future research to investigate and to distinguish between each sub-class of automaton. We formed our taxonomy according to the different modes of text production. Our efforts were primarily focused in separating any form of automation from organic, human content. In doing so we recognized three distinct classes of these types of automated accounts. However, boundary cases (e.g. cyborg-spammers, robot-spammers, robotic-cyborgs, etc.) along with other potential aforementioned subclasses (e.g. listers, quoters, etc.) can limit the prowess of our current classification scheme tailored towards these subclasses. We have shown that human content is distinctly different from these forms of automation, and that for a binary classification of automated or human, these features have a very reasonable performance with our proposed algorithm.

Our study distinguishes itself by focusing on automated behavior that is tolerated by Twitter, since both types of inorganic content can skew the results of sociolinguistic analyses. This is particularly important, since Twitter has become a possible outlet for

health economics [4] research including monitoring patient satisfaction and modeling disease spread [25,3]. Monitoring excessive social media marketing of electronic nicotine delivery systems (also known as e-cigarettes), discussed in [3,26], makes classifying organic and automated activity relevant for research that can benefit policy-makers regarding public health agendas. Isolating organic content on Twitter can help dampen noisy data-sets and is pertinent for research involving social media data and other linguistic data sources where a mixture of humans and automatons exist.

In health care, a cardinal problem with the use of electronic medical records is their lack of interoperability. This is compounded by a lack of standardization and use of data dictionaries which results in a lack of precision concerning our ability to collate signs, symptoms, and diagnoses. The use of millions or billions of tweets concerning a given symptom or diagnosis might help to improve that precision. But it would be a major setback if the insertion of data tweeted from automatons would obscure useful interpretation of such data. We hope that the approaches we have outlined in the present manuscript will help alleviate such problems.

Acknowledgments

The authors wish to acknowledge the Vermont Advanced Computing Core which provided High Performance Computing resources contributing to the research results. EMC and JRW was supported by the UVM Complex Systems Center, PSD was supported by NSF Career Award # 0846668. CMD and PSD were also supported by a grant from the MITRE Corporation and NSF grant #1447634. CJ is supported in part by the National Institute of Health (NIH) Research wards R01DA014028 & R01HD075669, and by the Center of Biomedical Research Excellence Award P20GM103644 from the National Institute of General Medical Sciences.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jocs.2015.11.002.

References

[1] V.S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, F. Menczer, arXiv:1601.05140, (2016), URL http://arxiv.org/pdf/1601.05140v1.pdf.
[2] D. Harris, Can evil data scientists fool us all with the world's best spam? 2013 goo.gl/psEguf.
[3] E.M. Clark, C. Jones, J.R. Williams, A.N. Kurti, M.C. Norotsky, C.M. Danforth, P.S. Dodds, arXiv:1508.01843, (2015), URL http://arxiv.org/abs/1508.01843.
[4] A. Sadilek, H.A. Kautz, V. Silenzio, ICWSM, 2012.
[5] A. Wagstaff, A.J. Culyer, J. Health Econ. 31 (2012) 406.
[6] L. Mitchell, M.R. Frank, K.D. Harris, P.S. Dodds, C.M. Danforth, PLOS ONE 8 (2013) e64417.
[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, CoRR abs/1407.5225, 2014, URL http://arxiv.org/abs/1407.5225.
[8] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Collaboration Electronic messaging, Anti-Abuse and Spam Conference (CEAS, 2010.
[9] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Proceedings of the 26th Annual Computer Security Applications Conference (ACM, New York, NY, USA, 2010), ACSAC'10, 2010, pp. 21–30, http://dx.doi.org/10.1145/1920261.1920265, ISBN 978-1-4503-0133-6.
[10] C.M. Zhang, V. Paxson, Proceedings of the 12th International Conference on Passive and Active Measurement, PAM'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 102–111 http://dl.acm.org/citation.cfm?id=1987510.1987521, ISBN 978-3-642-19259-3.
[11] K. Thomas, C. Grier, D. Song, V. Paxson, Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, IMC'11, ACM, New York, NY, USA, 2011, pp. 243–258, http://dx.doi.org/10.1145/2068816.2068840, ISBN 978-1-4503-1013-0.
[12] K. Lee, B.D. Eoff, J. Caverlee, AAAI Intl Conference on Weblogs and Social Media (ICWSM), 2011.
[13] K. Lee, B.D. Eoff, J. Caverlee, in: W.W. Cohen, S. Gosling (Eds.), ICWSM, The AAAI Press, 2010, URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2010.html#LeeEC10.
[14] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, B.Y. Zhao, Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, IMC'10, ACM, New York, NY, USA, 2010, pp. 35–47, http://dx.doi.org/10.1145/1879141.1879147, ISBN 978-1-4503-0483-2.
[15] J. Huang, R. Kornfield, G. Szczypka, S.L. Emery, Tobacco control 23 (2014) iii26.
[16] K. Thomas, C. Grier, V. Paxson, Presented as part of the 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (USENIX, Berkeley, CA 2012), 2012, URL https://www.usenix.org/conference/leet12/adapting-social-spam-infrastructure-political-censorship.
[17] X. Wu, Z. Feng, W. Fan, J. Gao, Y. Yu, Machine Learning and Knowledge Discovery in Databases, in: H. Blockeel, K. Kersting, S. Nijssen, F. Elezn (Eds.), in: Lecture Notes in Computer Science, vol. 8190, Springer, Berlin, Heidelberg, 2013, pp. 483–498, http://dx.doi.org/10.1007/978-3-642-40994-3_31, ISBN 978-3-642-40993-6.
[18] J.R. Williams, J.P. Bagrow, C.M. Danforth, P.S. Dodds, Text mixing shapes the anatomy of rank-frequency distributions: a modern Zipfian mechanics for natural language Phys. Rev. E 91 (5) (2015) 052811.
[19] S. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans. Depend. Secure Comput. 9 (6) (2012) 811–824.
[20] J.P. Dickerson, V. Kagan, H. Wang, V.S. Subrahmanian, 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, pp. 620–627.
[21] G. Brown, T. Howe, M. Ihbe, A. Prakash, K. Borders, Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW'08, ACM, New York, NY, USA, 2008, pp. 403–412, http://dx.doi.org/10.1145/1460563.1460628, ISBN 978-1-60558-007-4.
[22] C. Wagner, S. Mitter, M. Strohmaier, C. Körner, When social bots attack: Modeling susceptibility of users in online social networks, Making Sense of Microposts (# MSM2012) (2012) 2.
[23] K. Lee, J. Caverlee, S. Webb, Proceedings of the 19th International Conference on World Wide Web, WWW'10, ACM, New York, NY, USA, 2010, pp. 1139–1140, http://dx.doi.org/10.1145/1772690.1772843, ISBN 978-1-60558-799-8.
[24] K. Lee, J. Caverlee, S. Webb, Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'10, ACM, New York, NY, USA, 2010, pp. 435–442, http://dx.doi.org/10.1145/1835449.1835522, ISBN 978-1-4503-0153-4.
[25] D.A. Broniatowski, M.J. Paul, M. Dredze, PLOS ONE 8 (2013) e83672.
[26] J. Huang, R. Kornfield, G. Szczypka, S.L. Emery, Tobacco Control 23 (2014) iii26.

Eric M. Clark is an applied mathematician who is interested in implementing mathematical theory to solve real-world, interdisciplinary problems. His interests include (but are not limited to) Network Theory, Social Contagion, Computational Linguistics, Natural Language Processing, Machine Learning, Evolutionary Algorithms, and Complex Systems. Currently, he is working with the Department of Surgery at UVM to investigate health trends on Twitter to make socio-geographical comparisons of different treatment regimens and how sentiments surrounding such health disparities are changing over time.

Jake Ryland Williams Upon receiving his Ph.D. in mathematical sciences at the University of Vermont (UVM) spring, 2015, he accepted a postdoctoral fellowship at the University of California Berkeley with the School of Information in the Master of Information and Data Science Program, where he will teach coursework on machine learning and continue to explore his research interests at the intersection of mathematics, physics, and linguistics. While at UVM in the Department of Mathematics and Statistics and the Vermont Complex Systems Center, his thesis and focus of research was centered on mathematical linguistics and computational social science.

Chris Jones is a Health Economist, Assistant Professor and Director of the Global Health Economics Unit of the Center for Clinical and Translational Science, University of Vermont (UVM) College of Medicine. Prior to joining UVM, Dr. Jones worked in industry, as research faculty at the Johns Hopkins Bloomberg School of Public Health and as a government health economist in the National Institute for Health and Clinical Excellence (NICE) Collaborating Centre on Mental Health in London where he contributed to six U.K. national guidelines for the National Guideline Development Group, including the Guidelines for Preventing Physical Inactivity, gaining considerable expertise with health economic evaluation and health economic appraisal. Dr. Jones' work focuses on time-criticality, the cost-effectiveness of incentive-based treatments and patient centeredness. He has published novel

statistical methodologies for predicting cost and personalizing treatment pathways. He currently serves as an elected member of the New England Comparative Effectiveness Public Advisory Council and contributes to Vermont's State Blueprint for Health Analytic Working Group.

Richard A. Galbraith is a Professor in the Department of Medicine and also serves as the Director of the University of Vermont's Center for Clinical and Translational Science. The Center is dedicated to the concept of applying interdisciplinary research to translational science both from the bench to the bedside, from the bedside to the community, and from the community to health policy. He received his MD degree from King's College University, London England where he completed an internship and residency in Internal Medicine. He relocated to the United States and completed a fellowship in Endocrinology, Metabolism and Nutrition and earned an interdisciplinary PhD in Molecular and Cellular Physiology from the Medical University of South Carolina. He spent 12 years at the Rockefeller University Hospital in Manhattan, New York where he directed the Rockefeller University Hospital and its attendant translational research programs and was Attending Physician at the New York Hospital. He also served as the Medical Director and Hospital Administrator for six years.

Chris Danforth is the Flint Professor of Mathematical, Natural, and Technical Sciences at the University of Vermont. He co-directs the Computational Story Lab, a group of applied mathematicians at the undergraduate, masters, PhD, and postdoctoral level working on large-scale, system problems in many fields including sociology, nonlinear dynamics, networks, ecology, and physics. His research has been covered by the New York Times, Science Magazine, and the BBC among others. Descriptions of his projects are available at his website: http://uvm.edu/~cdanfort

Peter Sheridan Dodds is a Professor at the University of Vermont (UVM) working on system-level problems in many fields, ranging from sociology to physics. He is Director of UVM's Complex Systems Center, co-Director of UVM's Computational Story Lab, a visiting faculty fellow at the Vermont Advanced Computing Core, and is appointed to the Department of Mathematics and Statistics. He maintains general research and teaching interests in complex systems and networks with a current focus on sociotechnical and psychological phenomena including collective emotional states, contagion, language, and stories. His methods encompass large-scale data collection and analysis, large-scale sociotechnical experiments, and the formulation, analysis, and simulation of theoretical models. Dodds's training is in theoretical physics, mathematics, and electrical engineering with extensive formal postdoctoral and research experience in the social sciences. Dodds's foundational funding was an NSF CAREER grant awarded by the Social and Economic Sciences Directorate.