The Perception of Social Bots by Human and Machine

Scott Appling
(scott.appling@gtri.gatech.edu)
Georgia Tech Research Institute,
Georgia Institute of Technology
Atlanta, GA 30332, USA

Erica J. Briscoe
(erica.briscoe@gtri.gatech.edu)
Georgia Tech Research Institute,
Georgia Institute of Technology
Atlanta, GA 30332, USA

Abstract

In this work, we investigate how humans and machine learning algorithms may detect social (influence) bots (as opposed to human created ones). We concentrate on two primary questions of interest: (1) What features do humans attend to in order to decide whether a social media account is human?; (2) How do those features compare with those most useful in classification algorithms trained to identify social bot accounts? Our study uses a dataset that was collected as a part of a social bot discovery competition. We discuss the predictive value of a variety of structural- and content-derived features and how these qualities may be utilized to inform social bot detectors that operate optimally, relevant to both human- and machine-based perception.

Keywords: social bots; social media; bot detection; human perception of bots

Introduction

Since the rise of the use of computers, artificial intelligence researchers have been interested in creating seemingly-real computer-based artificial behavior, often focused on holding human-like conversation (as outlined in the Turing test (Turing 1950)). These conversation programs (or 'chat bots') are aimed at simulating natural language use. Perhaps the most well known of these conversation programs was the Eliza program (Weizenbaum 1966) which used a set of all-purpose phrases (e.g., How do you feel about that?) mimicking Freudian-style psychoanalysis.

With the rise in popularity of social media, there has been a similar increase in the number of automated programs that aim to replicate human behavior in this space (Hwang, Pearce, and Nanis 2012). These 'social bots', or virtual software agents, may be used to produce automated posts for a variety of reasons (Boshmaf et al. 2013), such as to produce both targeted and non-targeted promotional content (i.e., 'spam') or to produce support for a political candidate (Ratkiewicz et al. 2011). Social bots merely copy real (human) users' actions or function by employing artificial intelligence algorithms (Boshmaf et al. 2013).

Chat bots have been created using various natural language generation techniques. One of the most widely used methods is AIML (Artificial Intelligence Markup Language), which uses XML as a knowledge base along with pattern categories and response templates (e.g., (Mikic et al. 2009). Most of these approaches do not operate over historical conversations, responding only to the last post that was made. Advances in natural language generation, in addition to the increasingly sophisticated methods for detecting social bots, has led to a large number of fake accounts in social media platforms (e.g., (Gupta et al. 2013)).

Recently, much effort has been focused on the automatic detection of fake accounts, both by academics (e.g., (Dickerson, Kagan, and Subrahmanian 2014)) and the social media platform providers ((Twitter 2016a)). Methods for bot detection rely on machine learning to create models that utilize features available in the social media platform (e.g., (Chu et al. 2012)). Related work has also investigated the susceptibility of users to bot attacks, where users who exhibit more 'openness' are more likely interact with bots (Wagner et al. 2012).

Given these efforts, we are interested in exploring the efficacy of human detection of social bots and how that compares to machine-based detection. With the preponderance of research focusing on automated modeling techniques for identifying social bots, we find that little has been done to understand the ways in which humans perceive and decide about the nature of social media accounts they encounter within their online social networks. Here, we investigate how humans determine the validity of social media (Twitter) accounts through their observation of account characteristics. We concentrate on two areas of interest: The results of these two analysis allow us to understand not only the differences between the optimal set of features used for classification (as determined through machine learning) and those used by humans, but also how bot detectors may be built to optimize human bot detection behavior.

Our study uses a dataset that was collected as a part of a social bot discovery competition (Subrahmanian et al. 2016). We discuss the predictive value of a variety of structural- and content-derived features and how these qualities may be utilized to inform social bot detectors.

The rest of the paper is organized as follows: First, we discuss the creation and use of automated systems that have been design to interact with users in social media (social bots), including the variety and efficacy of methods used to detect them. We also describe the social bot challenge that was held in 2015 that serves as the source of our social bot behavioral data. Next we discuss the cues and methods that people use to authenticate behavior that they encounter in social media. We then cover two analyses. In the first analysis we evaluate how well humans perform at social influence bot[1] detection, using data that was collected during a bot detection competition based on the Twitter platform. The second analysis uses classification algorithms paired with human solicited confidence scores related to the human or non-human nature of accounts to determine the most important features for detecting social bots.[2] We then discuss the results and the implications of increasing social bot sophistication.

Perceptions in Social Media

Social media provides a rich environment in which users can interact and form impressions of other users' qualities, such as credibility or competence (e.g., (Briscoe, Appling, and Hayes 2014)). These perceptions may be based on properties that derive from the social network (e.g., number of followers), the user's profile information (e.g., profile picture), or the content (e.g., tweets) that the user produces (Flanagin and Metzger 2000); (Spence et al. 2013).

Twitter is an often studied (e.g.,(Lenhart et al. 2010), (Kwak et al. 2010),(Cha et al. 2010)) online social networking and microblogging tool, first released in 2006. Twitters micro-blogging platform allows users to post and read messages (tweets) of 140 characters or less. Users may subscribe to other users' Twitter feeds, share posts (or 'retweet'), and acquire followers of their own. Hashtags, namely words or phrases prefixed with a '#' symbol, can group tweets by topic. #JeSuisParis and #Jobs are two of the of the top two trending hashtags on Twitter in 2015 (Twitter 2016b).

For 'real' (human) users, creating a social media profile is fairly straightforward, requiring that they provide basic personal information (e.g., gender, location, etc.). For social bot accounts, the task is less clear, as automatic methods may not create realistic looking information, though bot designers may leverage work that has identified those qualities that are most socially desirable (Bilge et al. 2009). This information may be mined from other accounts or chosen through a random generation process.

Previous investigations have evaluated the link between Twitter profile features and how users infer qualities of the account, such as its credibility. Edwards et al., (Edwards et al. 2013) found that a user's influence score (in this study, represented as a Klout score) affected perceptions of that users credibility. Other cues that have been show to influence the perception of credibility include user name, number of followers, posted links leading to credible sites, other tweets communicating similar information, number of retweets, topical expertise, and reputation (Briscoe et al. 2013); (Morris et al. 2012). In a similar vein, we are interested in the features that subjects utilize to determine the human validity of an account in terms of it being human-run, as opposed to a social bot.

Bot Detection

Automated bot detection approaches often employ machine learning models that exploit some combination of social network structural and linguistic (including semantic and orthographic) features to predict likely bot accounts or humans who are likely to unknowingly interact with bots (Lee, Eoff, and Caverlee 2011; Wagner et al. 2012). These systems are typically trained in a supervised fashion on a set of known bot accounts as ground truth data. In (Lee, Eoff, and Caverlee 2011) a study was conducted to investigate the kinds of profile and content features that were exhibited by so-called "content polluters" towards their automated detection. The authors found that out of 16 features related to content, temporal changes in social network structure, and demographics, the top indicative features were related to the numerical IDs [3], as fake accounts were likely created near the same time. Similarly, other approaches have derived composed features, such as the follower-following ratio or distance between the victim and the spammer in the social graph (Thomas et al. 2011). These approaches, while initially effective, have been overcome by spam promoters that have created methods for obscuring or evading these metrics, such as through the creation of large bot networks (Amleshwaram et al. 2013).

The potential detrimental effect of rampant artificial, but realistic, social bots is evident. At the least, their saturation may make social media an unrealistic and unattractive virtual space. At worst, bots may have severe consequences, such as having negative effects on the stability of financial markets (Bollen, Mao, and Zeng 2011) or the proliferation of radical propaganda (Vidino and Hughes 2015).

In 2015, the Defense Advanced Research Projects Agency (DARPA) held a bot detection challenge that was specifically focused on the detection of 'social influence bots'. The data set was derived from a previous, independent competition held by Pacific Social Inc. in 2014, in which influence bots combated misinformation online, specifically around anti vaccine activists on Twitter. Social influence bots were created to intentionally participating in social media for the express purpose of influencing other users on a particular topic (here, centered on the topic of anti-vaccination) (see (Subrahmanian et al. 2016) for more details on the competition). The detection challenge was to create machine learning algorithms to determine the influence bots from the real accounts, which was performed offline on the datasets after the influence competition ended. We utilize this data set as it provides a unique opportunity to study how humans interact

[1] A social influence bot is a kind of social bot that is actively attempting to persuade users towards a particular topic and may change its approach based on interactions it has with individuals.

[2] In the first analysis we consider features as they are available to humans or algorithms whereas in the second analysis we focus solely on what features human can immediately observe as input for model building. For example, LIWC word category features are not immediately available to humans and not used in the second analysis.

[3] Twitter has in recent years replaced the account id assignment process with a non-sequential successor called *snowflake*

with sophisticated bots, specifically designed to maximize engagement and exert influence while avoiding detection.

Study: Comparing detection features between human and machine

The study was designed to determine the features that human use to determine non-human social media accounts and compare those features to those that prove most effective using a supervised learning approach.

Participants 209 participants were involved in the study. Participants were recruited and paid using Amazon's Mechanical Turk.

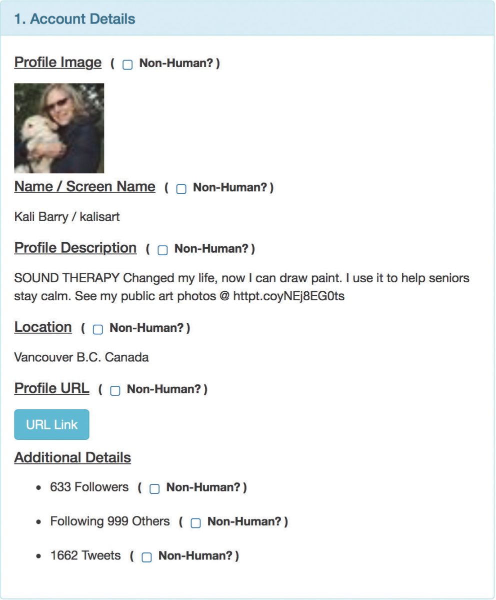
Method Our Twitter data was sampled from 3200 accounts that were accumulated as part of the DARPA Twitter bot challenge (see (Subrahmanian et al. 2016) for more details on the competition) that took place in 2015. From the larger set we selected the 39 accounts that represented the influence bots that were created and active during the competition. Additionally, we selected 39 other accounts to represent human (non-bot) accounts. These 'real' accounts were checked against the list of known bots that is published by Twitter, in order to ensure their human authenticity.

The following screen shot is an example of a profile that was shown to the human evaluators. The profile components that were included in the display were: profile image, screen name, profile description, location, profile URL, number of followers, number following, number of tweets posted, and the content of 10 randomly selected tweets from the account. One thing to note here is that while we did ask the subjects if they thought that the tweet content was realistic for a human writer, we did not explicitly ask them what features of that content were non-human like.

Subjects were instructed to look at the displayed account profiles and associated tweets and determine whether they thought the account was 'non-human' or 'human'. They were then asked to indicate for each particular feature (e.g., profile picture) whether the item looked like it was produced by a human or non-human. They were then asked to provide free form response explaining their decisions.

Analysis and Results On average, subjects predominantly thought most accounts were run by humans, where accounts were identified as bots 22% of the time.

A factorial logistic regression analysis revealed four statistically significant variables that were most predictive of an individual's decision on whether or not an account is human-based. These factors were: the profile description ($\beta = 0.759$, $p < 0.05$), screen name ($\beta = 1.175$, $p < 0.001$), the number of followers ($\beta = 1.479$, $p < 0.001$), and whether or not the tweet content seemed non-human ($\beta = 1.316$, $p < 0.001$).



1. Account Details

Profile Image (□ Non-Human?)

Name / Screen Name (□ Non-Human?)

Kali Barry / kalisart

Profile Description (□ Non-Human?)

SOUND THERAPY Changed my life, now I can draw paint. I use it to help seniors stay calm. See my public art photos @ httpt.coyNEj8EG0ts

Location (□ Non-Human?)

Vancouver B.C. Canada

Profile URL (□ Non-Human?)

Additional Details

- 633 Followers (□ Non-Human?)
- Following 999 Others (□ Non-Human?)
- 1662 Tweets (□ Non-Human?)

Figure 1: Example Twitter user profile that was shown to subjects, who were asked to rate the humanness of each of the features displayed.

Social Bot Detection Method

In order to understand the relative efficacy of the features that the human subjects used, we applied machine learning models that focus on choosing the account features most predictive of classifying an account as a social bot. In this case, we only utilize those features that were also available to the subjects, avoiding complicated metrics, such as temporal changes in ratios of followers to those following. The list of features was chosen to cover those that derive explicitly from the profile (e.g., profile description length), as well as linguistic cues calculated from the tweet content (e.g., negative emotion), using the validated categories provided by LIWC (Pennebaker, Francis, and Booth 2001). These linguistic features are intended to capture the qualities that the humans used in evaluating the tweet content.

We evaluated several machine learning algorithms including Linear Support Vector Machine (Fan et al. 2008), Random Forests (Breiman 2001), K-Nearest Neighbors (Altman 1992), and Extremely Randomized Trees (Geurts, Ernst, and Wehenkel 2006). We found that Extremely Randomized Trees gave the best performance. Table 1 lists the top 10 ranked features along with their corresponding feature weights. The higher the weight the more importance the feature plays in explaining the variance in the bot labels.

Building a Human-Based Bot Detector

The previously described study was a relatively small exploration of what features humans were using to decide between human and social influence bot accounts and how that compared the optimal set determined by a machine learning classification algorithm. To evaluate this behavior at a larger scale and to create detector based solely on features used by humans, we conducted an analysis over a larger twitter account dataset where humans were again asked to indicate the humanness of twitter account profile features and decide whether or not the account was non-human or human. In this

Table 1: The Top 10 Features used by Model

Feature	Weight	Description
Has Profile URL	0.105	Whether or not the profile had a URL link present.
Tweet Count	0.063	The number of status updates the account had.
Followers Count	0.061	The number of other accounts following the status update of the account.
Friends Count	0.051	The number of other accounts the account is following.
LIWC - 3rd Person Plural	0.036	The average percentage of words within tweets e.g. their, them, they, they'll
LIWC - Personal Pronouns	0.027	The average percentage of words within tweets e.g. she, I, me, my, he, we
Average Tweet Length	0.018	The average number of token present in the account's tweets.
Average Verb Count	0.025	The average percentage of verbs within tweets.
LIWC - Past Tense	0.022	The average percentage of past tense functional verbs within tweets e.g. did, didn't, could've, went, was
LIWC - Negative Emotion	0.018	The average percentage of negative emotion conveying words within tweets e.g. geek, weird, stupid

analysis, approximately 9600 human responses were collected, using Mechanical Turk, over 3200 twitter accounts (3 ratings for each account). In addition to deciding on whether or not an account was human or not, participants were also asked to indicate the confidence they had in their non-human or human designations where the estimates could range from 0 to 100.

Determining Feature Importance

To gain a sense of the importance of features that humans used in determining the nature of accounts, an extremely randomized trees regressor was trained with dichotomous features (i.e. is nonhuman, checked boxes) and scores formulated according to Equation 1. After training the regressor the feature weights were inspected and presented in Table 2. Accordingly, the most important feature for the model and what humans were placing much emphasis on, was the existence of any tweet content deemed non-human with the associated profile. This feature dominated with importance of 0.64; the next most important feature was the descriptive text in the profile with 0.21.

$$score = \begin{cases} 1 * confidence & is\_nonhuman = 1 \\ -1 * confidence & is\_nonhuman = 0 \end{cases} \quad (1)$$

Discussion

The purpose of the first study was to experimentally evaluate the features that people use to determine whether they are interacting with a bot in social media and the efficacy of those features as compared to a machine-trained classification algorithms. Similar to previous studies, where subjects found Twitterbots to be credible (Edwards et al. 2014), subjects were relatively poor at determining the validity of bot accounts (51% accurate), compared to 73% accuracy by the machine learning method.

The difficulty that the subjects had in identifying the bot accounts is likely due to a number of factors. First, identifying social bots in Twitter is made more challenging due to the fact that the communications are so succinct. This brevity obviously works in the botmaker's favor, as it minimizes the most challenging element to creating realistic social bots - the natural language generation component. Second, social bots are specifically designed to take advantage of social media platforms - capitalizing on their ability to adeptly construct and take advantage of social networks (Freitas et al. 2015). Third, humans are willing and ready to personify entities with whom they interact (Morris et al. 2012), making most social media users inherently vulnerable to social bots.

According to the detection study's regression results, the qualities that subject most utilized for their determination of a social bot were the "humanness" of the tweet, profile description, the name of the account, and the number of followers that the account had. Evaluation of the free form responses corroborates these results. Use of the profile description was indicative for both human and bot, where subjects reported "I think the name is not human. I also think that her profile description isn't human" and"This is a 100% verifiable human account. He has a well developed biography."

Most subjects commented on the content of the tweets (e.g., "No noticeable theme to the tweets that would make this seem like a real person with real interests" and "Many of the tweets seemed to be advertising something, suggesting the account may be a bot"). Though in our study, we did not ask the users to explicitly identify which aspects of the content aroused their suspicions, many of the features identified in the machine learning model were seemingly identified by the subjects. For example, the lack of personal pronouns (as identified by using LIWC) was one of the top ten predictive features in the classification model. This feature corresponds to the personal nature (or lack thereof) of the tweets, the presence of which makes them more human-like. A few comments did remark on this feature (e.g., "Personal opinion in multiple tweets led me to believe real person"). Another content feature that was often remarked on was the amount of negativity in the tweets (e.g., "Definitely a real person, he makes very angry posts."), which, in the machine learning model, was also found to be a discriminative feature.

Table 2: Feature Importance according to Human Perception

Feature	Weight	Description
Profile Image	0.032	The visual contents of the profile image.
Name and Screen Name	0.060	The given and surnames and twitter user name.
Description	0.211	The profile description text.
Location	0.008	The text the twitter user entered in the location area.
Profile URL	0.015	The URL link and the contents at the URL.
Followers Count	0.014	The number of other accounts following the status update of the account.
Friends Count	0.007	The number of other accounts the account is following.
Tweet Count	0.009	The number of status updates the account had.
Tweet Content	0.640	Randomly sampled tweet content associated with the profile.

Somewhat surprisingly, subjects also used the number of followers as an important quality (e.g., ”There are few followers, and the account follows few others”), which was also a discriminative feature as identified by the classification model. This is interesting as this quality is known to be widely manipulated (e.g., (Stevenson 2012)) and could be a direct result of the age of the account, where a new account would be expected to have few followers. Similar to studies on perceptions of credibility in Twitter (Morris et al. 2012), features such as profile photos were not consistently identified as indicative of humanness.

Useful qualities not identified as important by subjects, but were discriminative according to the classification model, included the presence of a URL in the profile and the number of previous tweets by the user. These qualities were mentioned in the free form responses (e.g., “The account only has one tweet so it is probably not human” and “There is only one tweet and that tweet is a completely human statement.”). Seemingly, subjects rationalized the number of previous tweets as either justifying or invalidating the user as being non-human.

The second study evaluated the optimal features for creating a machine learning classification algorithm that detects social bots as humans do. Given that the two most important features for this algorithm were text-based, in the future, we would likely build models that contain more nuanced language features.

A limitation of this study that should be acknowledged is the use of static screen shots to communicate the social media information. In reality, consumers of social media have long histories of interactions with their social network (both in real life and online), which likely affects how and what information they process when using a social media platform. Future studies may use longitudinal study designs or the subjects’ own networks to address this limitation. Additionally, we may use perceptual collection methods (e.g., eye-trackers) to understand how the users are attending to the presented information, instead of self-report, which may contain inherent biases.

Conclusion

Increasingly, people engage in social media interactions to communicate with one another, share knowledge, and to gain information. Because this medium allows for direct access to large groups of people, it provides the potential to engage with and influence large amounts of people with relatively little effort. This access to widespread influence is becoming more accessible through the use of sophisticated social bots. The results of our studies, in combination with others on credibility and bot detection in social media, provide the foundation for understanding the qualities most important for convincing a user that a bot is human. The manipulation of these qualities, by botmakers, could lead to increased deception; however, by understanding these qualities, social media platforms should be able to design detectors and interfaces that minimize the impact of fake accounts, both from a humanistic and machine point of view.

Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-12-1-0043.

References

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46(3):175–185.

Amleshwaram, A. A.; Reddy, N.; Yadav, S.; Gu, G.; and Yang, C. 2013. Cats: Characterizing automation of twitter spammers. In *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on*, 1–10. IEEE.

Bilge, L.; Strufe, T.; Balzarotti, D.; and Kirda, E. 2009. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web*, 551–560. ACM.

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8.

Boshmaf, Y.; Muslukhov, I.; Beznosov, K.; and Ripeanu, M. 2013. Design and analysis of a social botnet. *Computer Networks* 57(2):556–578.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Briscoe, E. J.; Appling, D. S.; and Hayes, H. 2014. Cues to deception in social media communications. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 1435–1443. IEEE.

Briscoe, E. J.; Appling, D. S.; Mappus IV, R. L.; and Hayes, H. 2013. Determining credibility from social network structure. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1418–1424. ACM.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, P. K. 2010. Measuring user influence in twitter: The million follower fallacy. *ICWSM* 10(10-17):30.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on* 9(6):811–824.

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

Edwards, C.; Spence, P. R.; Gentile, C. J.; Edwards, A.; and Edwards, A. 2013. How much klout do you have a test of system generated cues on source credibility. *Computers in Human Behavior* 29(5):A12–A16.

Edwards, C.; Edwards, A.; Spence, P. R.; and Shelton, A. K. 2014. Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior* 33:372–376.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9:1871–1874.

Flanagin, A. J., and Metzger, M. J. 2000. Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly* 77(3):515–540.

Freitas, C.; Benevenuto, F.; Ghosh, S.; and Veloso, A. 2015. Reverse engineering socialbot infiltration strategies in twitter. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 25–32. ACM.

Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. *Machine learning* 63(1):3–42.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web companion*, 729–736. International World Wide Web Conferences Steering Committee.

Hwang, T.; Pearce, I.; and Nanis, M. 2012. Socialbots: Voices from the fronts. *interactions* 19(2):38–45.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600. ACM.

Lee, K.; Eoff, B. D.; and Caverlee, J. 2011. Seven months with the devils: a long-term study of content polluters on twitter. In *AAAI Intl Conference on Weblogs and Social Media (ICWSM).*

Lenhart, A.; Purcell, K.; Smith, A.; and Zickuhr, K. 2010. Social media & mobile internet use among teens and young adults. millennials. *Pew Internet & American Life Project*.

Mikic, F. A.; Burguillo, J. C.; Llamas, M.; Rodríguez, D. A.; and Rodríguez, E. 2009. Charlie: An aiml-based chatterbot which works as an interface among ines and humans. In *EAEEIE Annual Conference, 2009*, 1–6. IEEE.

Morris, M. R.; Counts, S.; Roseway, A.; Hoff, A.; and Schwarz, J. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 441–450. ACM.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and tracking political abuse in social media. In *5th International AAAI Conference on Weblogs and Social Media*.

Spence, P. R.; Lachlan, K. A.; Spates, S. A.; and Lin, X. 2013. Intercultural differences in responses to health messages on social media from spokespeople with varying levels of ethnic identity. *Computers in Human Behavior* 29(3):1255–1259.

Stevenson, S. 2012. I bought 27,000 twitter followers. http://www.slate.com/articles/technology/technology/2012/10/buying_twitter_followers_is_it_worth_it_.html.

Subrahmanian, V.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F.; et al. 2016. The darpa twitter bot challenge. *arXiv preprint arXiv:1601.05140*.

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 243–258. ACM.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 433–460.

Twitter. 2016a. Fighting spam with botmaker. https://blog.twitter.com/2014/fighting-spam-with-botmaker.

Twitter. 2016b. Twitter top trends 2015.

Vidino, L., and Hughes, S. 2015. Isis in america: From retweets to raqqa (george washington university: Program on extremism, 2015) 7.

Wagner, C.; Mitter, S.; Körner, C.; and Strohmaier, M. 2012. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)* 2.

Weizenbaum, J. 1966. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.