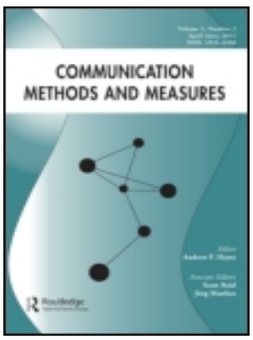


Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: http://www.tandfonline.com/loi/hcms20

Disentangling User Samples: A Supervised Machine Learning Approach to Proxy-population Mismatch in Twitter Research

K. Hazel Kwon, J. Hunter Priniski & Monica Chadha

To cite this article: K. Hazel Kwon, J. Hunter Priniski & Monica Chadha (2018): Disentangling User Samples: A Supervised Machine Learning Approach to Proxy-population Mismatch in Twitter Research, Communication Methods and Measures, DOI: 10.1080/19312458.2018.1430755

To link to this article: https://doi.org/10.1080/19312458.2018.1430755

Published online: 15 Feb 2018.

COMMUNICATION METHODS AND MEASURES
https://doi.org/10.1080/19312458.2018.1430755

Routledge
Taylor & Francis Group

Disentangling User Samples: A Supervised Machine Learning Approach to Proxy-population Mismatch in Twitter Research

K. Hazel Kwon [a], J. Hunter Priniski[b], and Monica Chadha[a]

[a]Walter Cronkite School of Journalism and Mass Communication, Arizona State University, Phoenix, AZ, USA; [b]Department of Mathematical and Statistical Sciences, Arizona State University, Phoenix, AZ, USA

ABSTRACT

This study addresses the issue of sampling biases in social media data-driven communication research. The authors demonstrate how supervised machine learning could reduce Twitter sampling bias induced from "proxy-population mismatch". Particularly, this study used the Random Forest (RF) classifier to disentangle tweet samples representative of general publics' activities from non-general—or institutional—activities. By applying RF classifier models to Twitter data sets relevant to four news events and a randomly pooled dataset, the study finds systematic differences between general user samples and institutional user samples in their messaging patterns. This article calls for disentangling Twitter user samples when ordinary user behaviors are the focus of research. It also builds on the development of machine learning modeling in the context of communication research.

"Just because Big Data presents us with large quantities of data does not mean that methodological issues are no longer relevant. Understanding sample, for example, is more important now than ever" (boyd & Crawford, 2012, p. 668).

Data-driven research has gained unprecedented popularity among communication scholars in the past few years. As the recent establishment of a Computational Methods interest group and the preconference workshop on computational techniques tools at the International Communication Association (ICA) conference in 2017 demonstrate (http://www.icahdq.org), communication scholarship has kept pace with the recent advancement of the computational social science paradigm. Social media platform Twitter, especially, has become one of the most popular data feeds for computational communication researchers who pursue "unobtrusive methods," (Burrows & Savage, 2014, p.2) to understand public opinions and behavioral tendencies of social media users in various social and political contexts. Indeed, social media research has offered fruitful insights on digitally mediated communicative practices, and new directions for research and theory-building.

Simultaneously however, data-driven research has raised concerns about its sampling procedure and whether it accurately represents the population or topic it claims to study. Critics have pointed out that existing computational models and their findings can not only pose biases toward issue experts, professionals, dedicated users, or certain demographic groups but also have low accuracy when they claim to represent the general body of "ordinary" users (Boyd & Crawford, 2012; Cohen & Ruth, 2013; Hargittai, 2015). In the same vein, communication research that examines large-scale data often does not disentangle the samples representative of "ordinary" publics from those of "specialized" actors. Rather, different types of users are mixed in an aggregated data structure that

CONTACT K. Hazel Kwon khkwon@asu.edu Walter Cronkite School of Journalism and Mass Communication, 555 Central Ave., Suite 302, Phoenix AZ 85004-1248.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hcms.

would treat them as a single population. Although studies have extensively discussed the roles of diverse social actors (e.g., elite vs. non-elite users) in shaping discourses in social media space (e.g., Himelboim, McCreery, & Smith, 2013; Papacharissi, 2015), a systematic procedure to unravel different types of social actors from the large data chunk has rarely been implemented into the sampling stage of such studies.

The current study extends the discussion on sampling biases in computational communication research. The underlying motivation of this study is to examine a machine-learning solution to improve sampling techniques for Twitter research, one of the most popular data venues for computational communication scholarship. Twitter sampling has introduced a variety of selection biases, caused by different issues such as proprietary restrictions (Burrows & Savage, 2014; Driscoll & Walker, 2014), algorithmic contamination (González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014; Salganik, 2017), noise from non-organic activities such as bots and astroturfing (Chu, Gianvecchio, Wang, & Jajodia, 2012; Ratkiewicz et al., 2011), and observational errors (boyd & Crawford, 2012; Hargittai, 2015; Salganik, 2017). Whereas some of these issues are beyond the individual researchers' control, others may be amenable by implementing rigorous sampling procedures and data preprocessing.

The purpose of this article is a methodological contribution, and it is organized as follows. In the following section, we discuss various social media sampling issues based on three concerns: predefined sampling frame, engineered noises, and proxy misspecification. After introducing the three concerns, the study narrows the focus down onto one specific issue related to proxy misspecification, the so called "proxy-population mismatch." It also is referred to as the "[unverified] quantitative relation between the proxy and the original populations studied" (Ruths & Pfeffer, 2014, p. 1064). Next, we introduce the research design in which we demonstrate the use of supervised machine learning as one approach to mitigate the proxy- population mismatch issue. As an example, we apply Random Forest classification models to separate ordinary users—whom communication scholars traditionally envision as constituents of the general publics and are referred to as *general public users* in this article—from the specialized actors who altogether we classify as *institutional users*. After explaining how we developed the model, we empirically examine whether the tweet sample representative of the general publics is indeed systematically different from the sample that represents institutional users. Based on the results, the conclusion section reiterates the need to couple data-preprocessing with systematic "slicing" of the dataset (Woodford, Walker, & Paul, 2013) to avoid confounding errors induced from proxy-population mismatch.

Background

Biases in social media data-driven research

Sampling of social media data, unfortunately, has not always been transparent. Literature has discussed various sources for biases in social media data-driven research, which we recap into three concerns: (1) predefined sampling frame, (2) noise from engineered activities, and (3) proxy misspecification. Note these three problems may not cover the exhaustive list of sources of biases and are not necessarily exclusive of one another either. Regardless, distinguishing these problems is helpful to elucidate sources of biases that can be improved through more systematic sampling or preprocessing procedures.

Predefined sampling frame. The first type of bias is induced from the use of a predefined sampling frame. Conventional social research, including communication studies, begins by setting up the research purpose and defining the "target population" (Salganik, 2017). Choosing the sampling frame—a pool of potential samples representative of the population—usually comes next. In social media data-driven research, however, the sampling frame is often predefined irrespective of the research purpose.

A primary reason for computational research using a predefined sampling frame is data inaccessibility, often attributed to the *proprietary restrictions* imposed by social media companies. Researchers may investigate a certain platform as an empirical site not always because the selected platform is the best representation of the phenomena under investigation but rather it allows easier access to data than other platforms (Schroeder, 2014). Twitter, for example, has undeniably become the most popular data source than other social media platforms for computational communication studies. This is not only due to Twitter's role in facilitating public communication but also the "relatively easy access to the data" enabled by the platform's application programing interfaces (API) policy (Gonzalez-Balion et al., 2014, p. 16). Conversely, private communication in more restrictive platforms such as Facebook or Snapchat is not accessible to majority of researchers, and thus precluded from the sampling frame regardless of the empirical relevance of such data to the research purpose or target population.

Furthermore, the same company's data policy varies from time to time, making the replication of a previous sampling procedure implausible even within a single platform context. This problem is referred to as "*system drift*" (Salganik, 2017). For example, studies that used ego network samples from Facebook (e.g., Brooks, Hogan, Ellison, Lampe, & Vitak, 2014; Kwon, Stefanone, & Barnett, 2014) may not be replicable anymore due to the companies' restrictions on the Application Program Interface (API) that previously enabled the collection of network data. Similarly, access to Twitter data has become increasingly limited over time. Unless data are purchased via the commercial "firehose" service, the currently available Twitter public API allows access to only 1% of all the tweets generated, for which "the methods that Twitter employs. . .is unknown" (Morstatter, Pfeffer, Liu, & Carley, 2013, p. 1). Additionally, studies have attested to systematic differences in measurements among the different API-based datasets (Driscoll & Walker, 2014; Gonzalez-Bailon et al., 2014; Morstatter et al., 2013).

Engineered noises. The second type of sampling bias arises due to the artificially planned activities that affect data creation or distribution, which together, we refer to as engineered noises. Engineered noises occur mainly from two factors: First, "*algorithmic confounding*" or "the ways that the goals of system designers can introduce patterns into data" (Salganik, 2017). For example, the distribution of a friend's network size on Facebook shows an anomalous peak at around 20 friends (Ugander, Karrer, Backstrom, & Marlow, 2011). Scholars familiar with the "social brain hypothesis" would interpret this anomaly as a spillover of the "sympathy group"–a grouping of maximum 20 individuals with whom one can maintain meaningful relations and contact on a regular basis (Zhou, Sornette, Hill, & Dunbar, 2005, p. 440)–on Facebook. In fact, this anomaly arises from the company's algorithmic curation that encourages "low friend count individuals" to connect with more friends "until they reach 20 friends" (Ugander et al., 2011, p.3).

Second, *non-organic activities* such as the use of bots and astroturfing also produce engineered noises. Studies estimate that 7% of Facebook accounts, and 9–15% of Twitter accounts are composed of bot-accounts (Varol, Ferrara, Davis, Menczer, & Flammini, 2017). Bot-assisted activities can manipulate public opinion (Ratkiewicz et al., 2011), contaminate social sharing processes (Lee, Eoff, & Caverlee, 2011), and disrupt information propagation (Abokhodair, Yoo, & McDonald, 2015). Sampling biases introduced from nonorganic activities could mislead our interpretation of the bottom-up processes underlying online public discourse. The noise from non-organic activities could be especially problematic when the research goal is to explore spontaneously emergent, grass-root phenomena in social media.

Proxy misspecification. In addition to the issues of predefined sampling frame and engineered noises discussed above, proxy misspecification can create another layer of errors. Particularly, defining proxies can vary, contingent upon how researchers set the data collection parameters such as keywords, time windows, locations, and others. That is, *misspecified or insufficient search parameters* could result in misrepresentation of the population under investigation. For example, González-Bailón et al. (2014) compared three different data collection methods on Twitter during the Indignados movement in 2012: SearchAPI using six hashtag search keywords; StreamAPI using

the same six hashtag keywords; and StreamAPI using an extensive list of 70 keywords. Their network analysis results revealed that the effect of search parameters was "more prominent" than the API effect in creating biases in network statistics (González-Bailón et al., 2014, p. 24).

Finally, another important issue that falls in line with the problem of proxy misspecification is called "*proxy-population mismatch*", which is the focus of the current study. Proxy-population mismatch refers to a gap between the proxy and real populations due to lack of understanding of the nature of the proxy (Hargittai, 2015; Hargittai & Litt, 2011). The proxy-population mismatch has been explained as a methodological challenge, as Ruths and Pfeffer (2014) state:

"Social media research question defines a population of interest: e.g., voting preference among California university students. However, because human populations rarely self-label, proxy populations of users are commonly studied instead, for example, the set of all Facebook users who report attending a UC school. However, the quantitative relation between the proxy and original populations studied, typically, is unknown—a source of potentially serious bias" (p. 1064).

Cohen and Ruth (2013) tap into this issue by re-examining predictive models of political orientations on Twitter. They found that the prediction tasks were largely successful when the samples were drawn from the population of politicians or politically active users. However, the same approach revealed much poorer outcomes when "more normal, less politically vocal" users were under investigation (p. 98). Their finding suggests that analytic efforts to understand social media users' attributes can overestimate the attribute tendency due to the bias toward specialized users who would explicitly display the attribute markers. In other words, random sampling from 'all' available Twitter data pool could be misleading in some contexts because the assumption that all Twitter users should be the proxy of ordinary social media users has not been verified. Other scholars have similarly contended that the social media population is not a proxy for the general population and the former is biased toward certain demographics and topical interests, recommending social media data-driven studies to be explicit about the research scope and the potential proxy-population mismatch (Hargittai, 2015; Hargittai & Litt, 2011).

Summary. Data-driven research is not free from sampling and measurement challenges. Some of the challenges are entwined with the social media company's policies and decision-making that are often unknown to, or beyond the control of, academic researchers. Some other issues arise due to the social media platform's inherent characteristics or third-party interruptions. Researchers may only reflexively respond to such issues, for example, by acknowledging data limitations and explicating the scope of the research (Hargittai, 2015). Meanwhile, some other methodological issues could be addressed by developing a rigorous research design. Extensive domain knowledge and effortful data collection and preprocessing techniques could improve the representativeness of samples. For example, there have been recent endeavors to disambiguate bots from a large-scale data corpus (e.g., Abokhodair et al., 2015; Chu et al., 2012; Varol et al., 2017). Cohen and Ruth (2013) demonstrated the use of machine-assisted sampling to help reduce errors introduced not only by bot activities but also by other sources of proxy-population mismatch. This study taps into this proxy-population mismatch issue, specifically addressing the question whether the general publics' activities on Twitter can be captured in a manner that represents the target population more accurately.

An example of proxy-population mismatch: proxy for general public users in Twitter

Twitter is a multifarious information stream that blends various groups of users, ranging from institutions, organizations, elites, to ordinary, non-elite individuals (Hermida, 2010). Examining all publicly available Twitter data has been useful in manifesting such diversities in social media communication networks (e.g., Jackson & Foucault Welles, 2015; Papacharissi, 2015). However, bundling all types of users and their activities into a single 'sampling basket' may not be the ideal approach when the research objective is to understand behaviors of *general publics* –a collective of individual citizens distinct from media elites, government, or other institutional entities.

Table 1. Examples of journal of communication articles based on quantitative Twitter analyses.

Article	Target Population	Twitter Data Access	Sampling Parameters	Separated between ordinary and institutional users?
Colleoni et al. (2014)	Twitter users who engage in political discourses	Secondary	Political users identified by using machine learning techniques from the secondary large-scale Twitter data of 2009.	No
Emery et al. (2014)	Online viewers of the Tips campaign	Firehose	Publicly available tweets collected using 36 search keywords during the campaign period; Used machine learning techniques to disambiguate Tips campaign relevant tweets.	No
Murthy et al. (2015)	General Twitter users	StreamAPI	Globally available tweets at any given time during the summer of 2013	No
Shin and Thorson (2017)	Users who share fact-checking messages in Twitter	Firehose	Publicly available political tweets searched by the list of 427 keywords during the 2012 presidential election; Used 194 tweets posted by three fact-checking websites to identify users who re-tweeted or commented on these tweets.	No
Vargo et al. (2014)	Supporters of presidential candidates	StreamAPI	Tweets made by supporters of Obama and Romney, identified by sentiment analysis coupled with machine learning, and tweets made by media accounts	Partly Yes (separated tweets sent from a predefined list of media accounts)

Table 1 lists a few exemplary studies published in one of the flagship journals in the communication discipline (*Journal of Communication*) that quantitatively analyzed Twitter data to understand communication patterns of social media publics. The list is far from exhaustive; the articles are high-quality publications that offer insightful findings using innovative data sources and methodological approaches. The intention is thus not to discredit the value of these studies but to accentuate a couple of points resonant with the issue of proxy-population mismatch.

First, the studies listed in Table 1 alluded to their population of interest as online "citizens" (Vaccari, Chadwick, & O'Loughlin, 2015), "viewers" (Emery, Szczypka, Abril, Kim, & Vera, 2014), "audiences" (Vargo, Guo, McCombs, & Shaw, 2014), "[information] consumer" (Shin & Thorson, 2017), or simply, "users" (Colleoni, Rozza, & Arvidsson, 2014). To represent these populations, they considered all publicly available Twitter data, searched by the researchers' filtering parameters. The actual data structure of Twitter, however, is likely to contain information produced by organizational, non-personal, or institutional activities. Therefore, assuming that all searched tweets represent a single population could neglect to account for a potential mismatch between the proxy—publicly accessible data in Twitter—and the population of interest—citizens, audiences, or ordinary users constituting the general public.

Second, some of these studies classified their data into subgroups, for example in terms of political orientations (Himelboim et al., 2013; Shin & Thorson, 2017; Vargo et al., 2014) or device types (Murthy, Bowman, Gross, & McGarry, 2015). The analytic focus then is to investigate differential effects of the sub-groups on shaping the patterns of tweet activities. Investigating subgroup effects has indeed resulted in shrewd lessons about social media use patterns. Nevertheless, the possibility of confounding effects due to the inherent differences within the proxies should not be neglected. For example, Shin and Thorson (2017) studied political fact-checking information diffusion, and underscored the role of "consumers' voluntary sharing and commenting behavior" in "increasing visibility of [fact-checking] messages," (p. 15). Meanwhile, less highlighted was the possibility that political fact-checking information could be disseminated not just by individual consumers' voluntary sharing but also by strategic,

coordinated political efforts. If the sample data contained a significant portion of strategic actors beyond the ordinary citizen population, the emphasis on voluntary sharing could be an overstatement. Another exemplary study (Murthy et al., 2015) showed the effect of devices (mobile-based vs. web-based) on shaping linguistic styles in tweets. In discussing their findings, however, Murthy et al. (2015) cautiously suggest the possibility of spurious inference, highlighting "the value of keeping the question open as to whether the type of people who tweet from mobile devices are qualitatively different from those who tweet from web-based platforms," (p. 834). That is, if ordinary users are likely to use mobile devices to send their tweets and institutional actors send their tweets via web platforms, or vice versa, the linguistic difference could be produced not due to device affordances but rather the distinct natures of populations—the general populace versus institutional actors.

To summarize, communication activities on Twitter are social and non-social, and spontaneous and strategic. When the focus of research lies on naturally occurring communication among ordinary individual users, the assumption that all publicly available Twitter data should be the proxy population could engender an issue of proxy-population mismatch, and lead to a misinterpretation of findings. To address this concern, the study explicated here explores the use of computerized processes to help researchers resolve the issue of proxy-population mismatch and thus proposes the first RQ.

RQ1: Could machine learning help disentangle the separate populations, specifically one representative of *general public users* from those comprising specialized user accounts such as media elites, government, and other institutional entities, organizational agents, and non-personal users?

In addition to effectively separating the two sample groups, we tested for differences in the tweeting patterns of general public users and institutional users. We conducted post-hoc tests with descriptive statistics of popularly studied characteristics of tweet messages, such as retweeting, the use of external informational source (URL), hash-tagging, and language uses (analytic and affective). The post-hoc tests intend to address whether there is indeed a possibility of sampling bias if these two groups are treated as one single population. Hence, the following hypotheses are proposed.

H1: The group of general users will show a different pattern of retweeting than institutional users.

H2: The group of general users will show a different pattern of using external information (URL) than institutional users following violent events.

H3: The group of general users will show a different pattern of hashtag use than institutional users.

H4: The group of general users will show a different pattern of analytic language use than institutional users following violent events.

H5: The group of general users will show a different pattern of affective language use than institutional users.

Additionally, we tested for differences in the tweeting patterns between the two populations based on the type of device used to tweet the message (Murthy et al., 2015). Only a couple of datasets include the related data field, and thus the following hypothesis is tested with the related datasets.

H6: The group of general users will show a different pattern than institutional users based on the device used to tweet the message.

Research design

Case studies: tweets on violence news

In this study, specialized actors have been grouped together and labeled as *institutional users*. For empirical exploration, we used four Twitter datasets that were collected after four violence-related news events: Mass shooting in Mesa, Arizona, a city in the metro-Phoenix area, in March 26, 2015 (Mesa), Boston Marathon Bombing in April 15, 2013 (Boston), Brussels Airport Attacks in March 22, 2016 (Brussels), and Quebec Mosque Attack in January 29, 2017 (Quebec).

One of the authors had previously collected the four Twitter datasets after the aforementioned news events as part of a larger research initiative. We used these news events as our empirical contexts for a few reasons. First, the episodic nature of such news tends to engender public responses within a short period time, and is, thus, relatively free from the temporal confounders. Second, terror news appeals to both, general publics and institutional actors, and thus may be advantageous for representing both populations. Third, violent incidents are usually breaking news events, that are less likely to involve planned activities than organized or longer-lasting news events such as political campaigns or social movements. Lastly, these four datasets represent different spatial scopes—local, national, and international—and, thus, the comparison of these datasets is opportune for exploring whether the machine-assisted sampling technique is applicable across different geographical scales.

The data for Mesa and Brussels were collected via NodeXL, a StreamAPI-based collection tool (Smith et al., 2010). Boston and Quebec data were collected by directly using StreamAPI. Mesa data contained tweets from March 18 (the day of the incident) through March 26, 2015, searched with the keywords *#MesaShooting*, and *'Mesa & Shooting'*; the coverage of Boston data was more sporadic, including April 15 (the day the bombing occurred), April 19–20, and April 24–April 29, 2013, searched with the keywords *#BostonMarathon, 'Boston & Marathon'*, and *Boston*. Brussels data was collected over a span of almost two weeks, from March 22 (the day of the bombing) to April 3, 2016; the keywords used in the tweet collection were *#BrusselsAttacks* and *Brussels*. Quebec data contained tweets from January 19 (the day of the shooting) to February 14, 2017, searched with the keywords *#QuebecAttacks, #QuebecShooting, 'Quebec & Mosque'*, and *Quebec*. To minimize cultural differences, we limited the investigation to English language tweets that originated from U.S.-based accounts. To identify the U.S.-based tweets, we developed a Twitter geo-tagging tool, using the four sources of geographic information associated with a tweet: tweet coordinates, tweet place, location in user profile, and tweet text, based on the "Carmen" library (Dredze, Paul, Bergsma, & Tran, 2013). As a result, the total number of tweets investigated was 7898, created by 4,345 unique users for Mesa, 14,211 tweets created by 9,562 unique users for Boston, 13,660 tweets created by 9,980 unique users for Brussels, and 151,928 tweets created by 87,225 unique users for Quebec. Note that these users are rarely single-tweet users in the datasets; the number of users who tweeted less than five times since the start date of their membership were only 14 users in Mesa, 26 users in Brussels, and 101 users in Quebec dataset (We did not have user status information for the Boston dataset).

Given that all data were drawn from news events similar in nature (i.e., violence news), we additionally created and analyzed a random dataset to enhance the generalizability of the results. This additional dataset included randomly pooled profile texts and tweets via StreamAPI for one week from September 22–28, 2017, at various times each day. Only tweets and profiles written in English from U.S.-based accounts were included (N = 175420).

For the post-hoc tests, retweet status, number of hashtags, and the use of URL were coded from the tweet messages. The use of analytic and affective languages was measured using Pennebaker, Booth, Boyd, & Francis, 2015 a tested lexicon-based sentiment analysis software (Pennebaker et al., 2015). Only two of our datasets (Boston and Quebec) contained information about tweeting devices and thus we looked at these two events for examining differences in the device popularity between the two populations. We replicated the process put forth by Murthy et al. (2015) to code the device types.

Technical description of random forest classification as a case for supervised machine learning in data preprocessing

Supervised Machine Learning modeling enables automatic classification of documents by feeding given algorithms with previously categorized data, also called the "training set". The algorithms can then apply the rules of the categorized data in the training set to uncategorized data in a "testing set". In this study, our goal was to classify two classes of Twitter users: general public users and institutional users. Specifically, given a training set of linguistic features extracted from Twitter profile descriptions, the Machine-Learning (ML) task was to approximate a function that successfully maps the linguistic features from the profile descriptions to the proper classes: general or institutional user. The model performance was then evaluated by applying the trained algorithm to a testing set, which would result in quality assurance metrics such as accuracy, precision, recall rates, and F-score.[1]

There are several popular classifier algorithms in Machine Learning. Random Forest (RF) classification is among the most popular and is the focus of this study. The RF classifier is a member of a class of learning algorithms known as Ensemble Learners that aggregate the results of many variating, 'weaker' learning functions to derive a final classification. Specifically, the RF classifier constructs multiple different decision trees that classify the data into the desired samples or categories and derives a final decision by aggregating all the trees' decisions. Hence, the model as a whole is analogously understood as a "forest," where many individual decision "trees" are grown. Based on the functions and trained dataset, each tree "votes" for a classification. The final decision or classification chosen by the forest is the one that receives the most votes. The RF classifier is particularly appropriate when the classes are not linearly separable. We assume that ordinary users are a heterogeneous collection of individuals, and thus our classification problem is not linearly separable (Figure 1). Although we chose the RF classifier for the prediction task given the linearly non-separable data condition, it does not mean that RF is superior to other algorithms. Accordingly, we present the results from other popular classifiers along with the RF results in the following section, that allow for an objective comparison of model performances across different classifiers.

In the RF classifier, each decision tree grows from a "root" node (no incoming edges, only outgoing edges), to "branch" nodes (has an incoming and an outgoing edge), and to "leaf or terminal" nodes (only has incoming edges). The process of splitting nodes is equivalent to the process of feature partitioning, resulting in one of the terminal nodes serving as the predicted class. In each tree, the nodal splits follow a random process, which is why the algorithm is called a *Random*

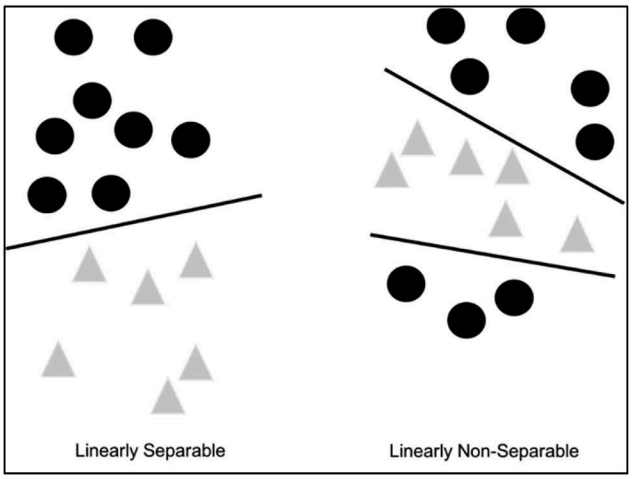


Figure 1. Example of linearly separable and non-separable data structures.

Forest (Breiman, 2001). Furthermore, each decision tree does not partition the data on the same set of features. Each tree is constructed from a random subset of features representing the data. The splitting or partitioning process continues recursively until the recursion reaches the best "split" that no longer makes the probability of a correct classification for the data much larger than the previous iteration (Loh, 2011). The decision on the "best" split is based on Gini impurity (GI)—the frequency measure for the inconsistent classification by a given feature set (Breiman, 1984).[2] Whereas the lowest Gini impurity would render perfect accuracy for group classification in an ideal world, the real data typically has an overfitting issue when the trees grow fully (unpruned) until each leaf node corresponds to the lowest impurity. Therefore, to avoid overfitting, a common practice is to "prune" trees, i.e., discontinue splitting once the decrease in GI from one split to the next is smaller than a given threshold (Breiman, 2001). Once individual trees make decisions after the partitioning, the forest derives the final prediction by aggregating the results using a method known as Bagging or **B**ootstrap **Agg**regat**ing** (Breiman, 2001). Figure 2 summarizes the RF classifier algorithm.

Classification of Twitter user samples

In each dataset (Mesa, Boston, Brussels, and Quebec), Twitter users were classified as either general public users or institutional users according to the five steps described below. After many trials and tests, this article presents the most successful procedure.

Step 1: rule-based annotation

In a supervised ML model, a machine learns from a training set and tests its learnt algorithms against a testing set. In this study, prepping the training and testing sets required the categorization of some profile texts as belonging to either general public or institutional users. Therefore, we randomly selected and labeled 1,000 user profile texts from the Mesa dataset, 2,000 from Boston, 2,000 from Brussels, and 2,000 from Quebec, a total of 7,000 texts.

As diverse and heterogeneous as general publics are, it could be daunting to consider every instance of ordinary user profiles. Instead, a more plausible approach was to adopt a process of elimination by which we labeled a profile as the positive class of the general user (=1) if it did *not* fall in any of the predefined non-ordinary user categories, namely: media organizations, journalism personnel, group, or organizational profiles, strictly professional career accounts with clear indication of organizational information, political/politicians account, and promotional account. The profiles falling in any of these categories were labeled as the negative class, which was the

Random Forest Classification Algorithm

Let our training set be $\boldsymbol{D} = ((X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n))$. (X_i, y_i) is a feature vector and its classification. $X_i \in X^d$, $y_i \in \{0,1\}$, where d is the number of features.

Construct T random decision trees

For $i = 1, 2, 3, \ldots, T$ decision trees:

Sample with replacement a $D_i = (X_i, y_i)$ from $\boldsymbol{D}$:

Uniformly and randomly select M features of X_i (**note:** $m << d$):

Classification tree $T_i \leftarrow$ **For** each $m \in M$: calculate **Gini**(m)

Aggregate results for prediction

Given a new $x' \notin \boldsymbol{D}$ (test-set)

Classification for $x' \leftarrow$ Majority vote of the B decision trees.

Figure 2. Summary of the RF classification algorithms.

institutional user (=0). Two graduate student coders annotated the selected profiles (Cohen's Kappa = .72), resolving disagreement through discussions when it occurred. The detailed description of this coding framework is found in our sister paper (Kwon, Chadha, & Pellizzaro, 2017).

Step 2: data preprocessing

The 7,000 labeled profile texts were pre-processed by (i) removing non-English and non-alphabetic characters, (ii) casting all letters in lower case, (iii) unifying all URLs into a single feature called "*http*", (iv) unifying Instagram links into a single feature "*ig*", and (v) removing stop-words based on the NLTK dictionary. We retained some stop-words like first-person pronouns, considering the possibility of a plural pronoun (such as "we") resulting in difference from a singular classifier (such as "I"). We did not employ stemming because the process led to lower accuracy results in all datasets. The labeled profile texts were then separated into two sets: a training set (80% of the total labeled data) and a testing set (20%), resulting in four training sets and four testing sets, each of which represented the corresponding Twitter dataset (Mesa, Boston, Brussels, and Quebec).

Step 3: feature extraction

The *sci-kit learn* package in Python was used to create the bag-of-words model, which turned each profile document into a feature vector. To avoid overfitting, we restricted—or "pruned"—the length of the feature vector to 500 most commonly used terms, which we chose based on term frequency (TF), by the raw count of a term. While we also tried to select 500 features based on term frequency-inverse document frequency (TF-IDF), the results remained largely the same. Previously, we tried to include bi-gram features and part-of-speech tagging; however, the model performance did not improve notably while the cost of computing time was greater. Accordingly, we decided to adhere to the simplest approach, which was to generate features based on TF. We then vectorized each profile text by representing it as a sparse (with mostly zeroes), 500-count frequency long vector.

Step 4: classification learning

The RF classifier was fed with each training set. For illustrative purpose, we take two examples of raw profile texts from the Mesa dataset. The first one was labeled as 'institutional user profile' (= 0), and the second one was labeled as "general public user profile" (=1).

Text 1 (institutional): Journalist and Reporter for AZ Republic.

Text 2 (general public): Love America. Bless God and go Cardinals! ig:instagram/ID

After preprocessing, each text would become: "*journalist reporter az republic*" and "*love america bless god go cardinals ig*". Let us assume that decision trees were constructed from the training set and the feature vectors used to partition the data were *journalist, ig, reporter, america, god*. Next, let us suppose that a text that included the features *journalist* or *reporter* was highly likely to be an institutional account, and conversely, a text containing the features *ig* (abbreviation for Instagram), *america*, or *god*, was more likely to belong to a general public user. Figure 3 presents a simplified example of the two decision trees concerned with these features for each text. These decision trees were iterated through recursive partitioning until they reached the best splits. The partial decisions derived from these decision trees were then aggregated for the final classification prediction. We used the *scikit-learn Random Forest module* in Python for the classification learning, with the number of trees set to 100.

Step 5: validation and automatic classification

After feeding the classifier with the training set data, a testing data set was used to validate the model's performance in accuracy, precision, and recall rates. The Results section below presents the RF-based validation results along with the results from other popular algorithms

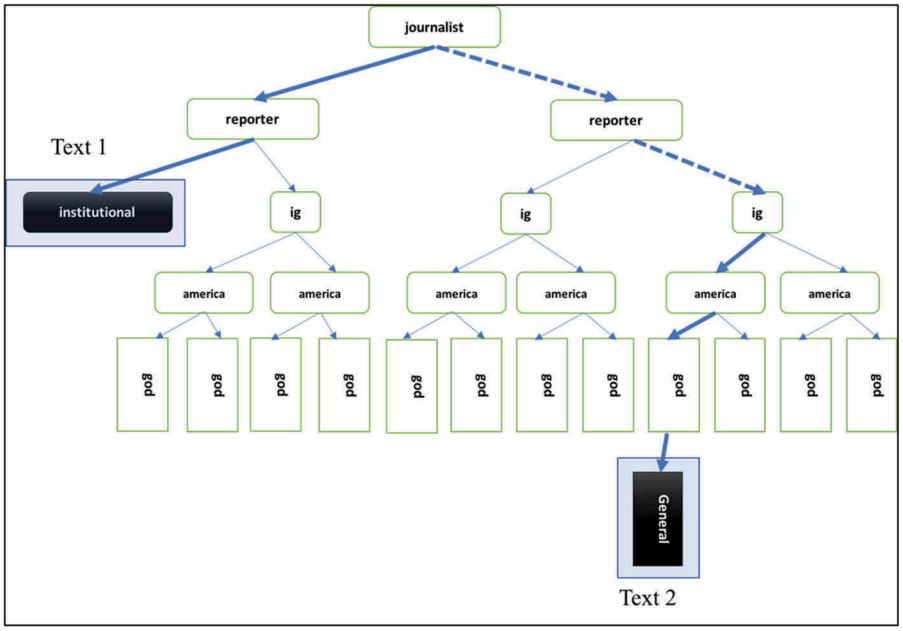


Figure 3. An example of two decision trees concerned with the features *journalist, ig, reporter, america, god* (a decision tree for text 1 = "journalist and reporter for az republic."; and another decision tree for text 2 = "love america. bless god and go cardinals! ig: instagram/id."). Bold line indicates the feature exists in the text; dotted line indicates the feature does not exist in the text. Each tree ends with a different class as the leaf node.

including SVM, Multinomial Naïve Bayes, and Logistic Regression. After the validation, the learnt RF algorithm was used to classify the rest of the unlabeled profiles as either general public users or institutional users. That is, two sample groups were created based on the machine-generated classification, and the differences between the two groups were examined using descriptive statistics.

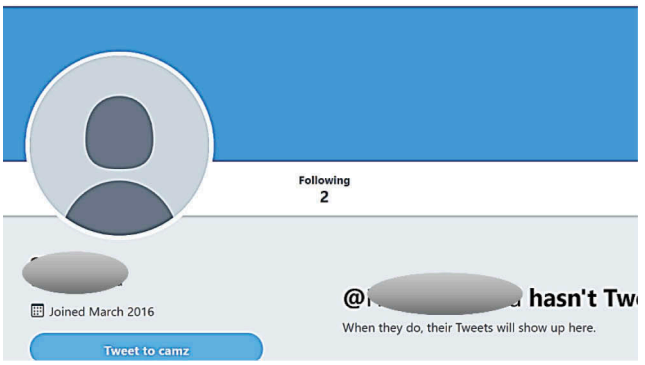
Results

Validation of human annotation outputs

As our approach for annotation was rule-based, we needed to validate the annotation outputs by manually reviewing actual user profile pages on Twitter where available. We could examine user profile pages for all datasets but Boston, for which we did not include the profile URLs field when collecting the data in 2013. The annotation outputs and manually checked labels matched highly, at 85.4% agreement. Some profiles (11%), however, were either suspended or did not exist any more, thus were impossible to validate. There were 3.6% unmatched profiles out of the profiles that were manually reviewed. The mismatch arose for several reasons, for example ambiguity about whether the account was a political promotional bot or belonged to a real human; amateur media/art/entertainment promotional accounts; freelance journalist profiles that had a journalistic profile description but the actual Twitter use was more personal; the accounts where tweets were either absent or had disappeared; a real person's account with no or little textual information in the profile description section. Figure 4 shows a few example profiles that showed inconsistency between the rule-based annotation and manual check of the accounts. Considering that the rule-based annotation resulted in only a small proportion of misclassification, we maintained our coding rules as a reasonable strategy for the purpose of this study.

(a) Annotated as institutional (i.e., promotional account); The actual profile was used primarily for personal tweets.

(b) Annotated as personal; The actual profile was used primarily for political promotion, giving suspect whether it is a bot.



(c) Previous had profile description annotated as a general user; The actual profile is unverifiable.

Figure 4. Examples of mismatched profiles (identifiable information shaded).

Validation of model performance

The hand-coded annotation resulted in 19.6% institutional users in Mesa data (N = 196 out of 1,000), 38.6% in Boston (N = 771 out of 2,000), 26.4% in Brussels (N = 527 out of 2,000), and 9.9% in Quebec (N = 527 out of 2,000); the machine-generated annotation included 16.4% in Mesa (N = 1296 out of 7,898), 32% in Boston (N = 4547 out of 14,211), 21.4% in Brussels (N = 2929 out of 13,660), and 7.2% in Quebec (N = 11,046 out of 151,928).

Next, the ML model performance was validated for each case, resulting in satisfactory recall rates ranging from 87.7–97.9%; precision rate ranging from 83.6–95.3%; accuracy ranging from 83–94.2%, and F1 score ranging from .86–.97. The RF results were similar to the results drawn from the other popular ML algorithms.

To examine the possibility of using labeled data in multiple news event contexts, we conducted cross-validations by using different news events' testing data sets. For example, the model trained with the Mesa training set was validated using the Boston, Brussels, and Quebec testing sets. Accordingly, nine additional pairs of training-testing-sets were examined for cross-validations. The overall results of cross-validations were high with the exception of the Boston dataset that provided lower scores for recall, precision, accuracy, and F1.

To assess generalizability, the same validation process was run with the Random dataset, comprising randomly selected tweets on uneventful days. Hand-coded random data included 18.1% institutional users (406 out of 2,247); the machine-based annotation included 17.68% (27,051 out of 152,983). Overall, the model performed slightly poorly with Random data, resulting in 74.3% for accuracy, 86.7% for precision, 82.7% for recall, and .85 for F1 measure. Cross-validations also suggest the model trained with Random data performed weakly in predicting classifications in other datasets. The results suggest an advantage of training the model with labeled data generated within a well-defined topical domain. Table 2 and Figure 5 presents the results for both within- and cross-validations of RF models, in comparison with other classifier results.

The list of feature coefficients quantifies the importance of features in predicting the user classes: a feature with a larger coefficient size plays a more important role in discriminating the two classes than one with a smaller coefficient. The results of RF coefficients, however, do not inform the features' directionality, i.e., which features are associated with the general user class, and which with the institutional user class. For this reason, we compared the RF-derived coefficients alongside the Logistic Regression (LR)-derived coefficients, which allow us to interpret the features' directionality. The list of top 50 feature coefficients is presented in Figure 6.

Table 2. Random Forest model validation results in comparison with other classifier algorithms. (Bolds = within data validation).

		Random Forest				Logistic Regression				Multinomial Naïve Bayes				Support Vector Machine			
Train	Test	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
Mesa	Mesa	**.942**	**.953**	**.979**	**.966**	.913	.922	.979	.950	.890	.950	.917	.933	.919	.946	.959	.952
	BMB	.655	.628	.989	.768	.675	.641	.992	.779	.763	.721	.962	.824	.692	.660	.966	.784
	Brussels	.774	.767	.983	.861	.789	.780	.983	.869	.808	.831	.920	.873	.794	.795	.958	.869
	Quebec	.887	.894	.989	.939	.887	.894	.989	.939	.846	.914	.911	.913	.864	.898	.954	.925
	Random	.802	.848	.933	.889	.833	.848	.978	.908	.765	.850	.878	.864	.777	.846	.900	.872
Boston	Boston	**.830**	**.836**	**.877**	**.856**	.843	.804	.962	.876	.814	.776	.954	.856	.841	.823	.923	.870
	Brussels	.736	.876	.736	.800	.796	.860	.854	.857	.836	.851	.934	.891	.759	.858	.795	.825
	Mesa	.850	.969	.848	.904	.843	.804	.962	.876	.836	.851	.934	.891	.850	.976	.841	.904
	Quebec	.806	.923	.851	.886	.851	.907	.926	.916	.846	.914	.911	.913	.793	.906	.854	.879
	Random	.685	.852	.760	.803	.789	.852	.909	.880	.733	.854	.825	.839	.757	.857	.856	.856
Brussels	Brussels	**.828**	**.874**	**.889**	**.881**	.861	.872	.944	.907	.838	.883	.892	.888	.826	.856	.910	.882
	Boston	.794	.762	.935	.840	.781	.737	.966	.836	.830	.801	.939	.864	.785	.744	.958	.838
	Mesa	.867	.942	.897	.919	.919	.940	.966	.952	.890	.970	.897	.932	.902	.957	.924	.940
	Quebec	.869	.919	.934	.926	.887	.911	.966	.938	.846	.921	.903	.912	.861	.913	.931	.922
	Random	.704	.859	.778	.817	.773	.847	.894	.870	.745	.857	.838	.848	.741	.849	.844	.847
Qubec	Quebec	**.889**	**.914**	**.966**	**.939**	.902	.906	.991	.947	.889	.923	.954	.938	.887	.916	.960	.937
	Brussels	.811	.827	.931	.876	.794	.784	.983	.872	.764	.780	.934	.850	.831	.824	.972	.892
	Boston	.770	.732	.950	.827	.699	.661	.985	.791	.748	.706	.966	.816	.737	.694	.973	.810
	Mesa	.879	.903	.959	.930	.890	.904	.972	.937	.873	.936	.910	.923	.867	.901	.945	.923
	Random	.792	.848	.919	.882	.843	.848	.991	.914	.791	.850	.914	.881	.820	.849	.958	.900
Random	Random	**.743**	**.867**	**.827**	**.847**	.835	.862	.962	.909	.762	.854	.871	.863	.778	.861	.885	.873
	Brussels	.616	.751	.714	.732	.712	.740	.940	.828	.683	.743	.870	.802	.695	.745	.891	.811
	Boston	.586	.636	.764	.694	.633	.632	.962	.763	.616	.636	.876	.737	.633	.638	.927	.756
	Mesa	.704	.863	.755	.806	.757	.826	.888	.856	.727	.846	.812	.829	.726	.825	.841	.833
	Quebec	.677	.897	.725	.802	.813	.904	.887	.895	.733	.903	.788	.842	.757	.902	.820	.859

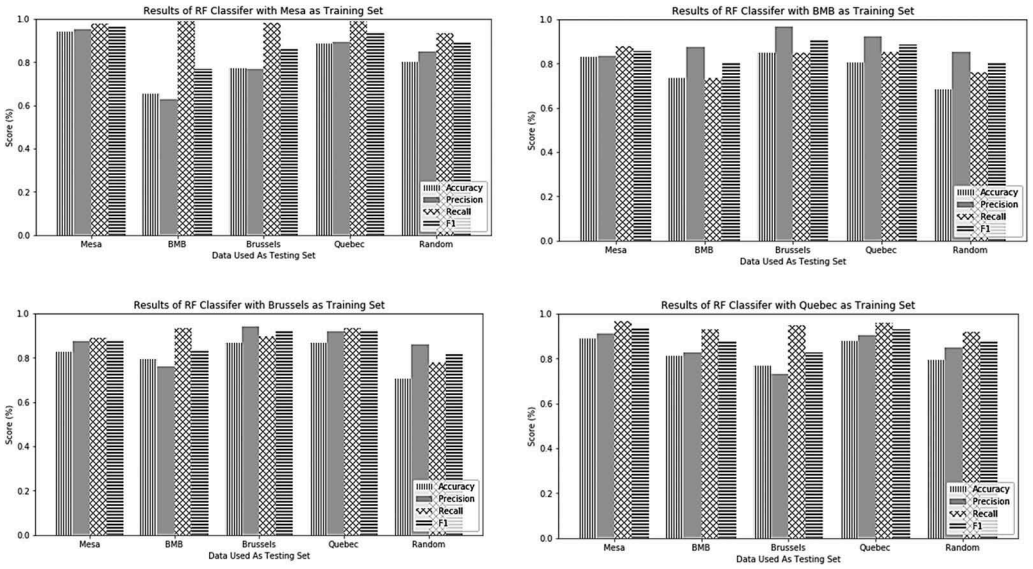


Figure 5. Visual summary of random forest model validation results.

The review of feature coefficients suggested that majority of high-ranked features were indicative of institutional users; and several of these features showed large coefficient sizes across all four datasets. For example, media channels (e.g., *channel, station, tv, network*) and journalism careers (e.g., *news, reporter, journalist, anchor*), and industry-related terms (e.g., *co, com, manager*) were ranked highly and associated with the institutional user class, with negatively signed coefficients. The presence of a hyperlink in the profile was also highly associated with the institutional user class (*http*). In the feature list for Mesa, locality-indicative words were ranked highly and associated with the institutional actor class, such as, *arizona, phoenix, valley, azcentral* (regional newspaper), *gilbert* (name of a town), and *kfyi* (local radio station). The plural first-person pronouns, such as *we* and *our*, were associated with the institutional user class, whereas singular pronouns *my* and *me* were associated with the general public user class. Some lifestyle related words such as *life, fan, music, love, student, artist*, or *american* showed stronger associations with the general public user class. Overall, the institutional user-related features were more prominent and consistent than general public-related features. Meanwhile, the top feature list for the Random dataset included fewer media-related terms than the other four datasets.

Error analysis

Our classification model is not error-free. A systematic analysis of the errors committed during testing should offer insight for future advancement on this model. For error analysis, we investigated two types of classification errors in each dataset, specifically False Positive classifications and False Negative classifications. Overall, the model performed more poorly in capturing negative classifications than positive classifications, missing true negative samples more frequently. The inability of the model in identifying true negative samples resulted in higher proportion of False Positive classification (Table 3).

Since the number of negative samples in our dataset is smaller than positive samples, our model sees less examples of organizational/institutional accounts in training. This leads to a small set of features that are representative of negative classifications. Since human coders can pick up on context clues and background semantic knowledge of features to derive classifications, this may not be a problem for humans. However, the model used in this research is unable to do this. To

MESA			BOSTON			BRUSSELS			QUEBEC			RANDOM		
Feature	RF-Coeff	LR-Coeff	Feature	RF-Coeff	LR-Coeff	Feature	RF-Coeff	LR-Coeff	Feature	RF	LR	Feature	RF	LR
news	0.0928	-2.351	news	0.1095	-2.472	news	0.0836	-2.024	journalist	0.0625	-3.3469	personally	0.0332	-1.0777
reporter	0.0432	-1.576	reporter	0.0319	-2.767	reporter	0.0273	-2.595	news	0.0440	-2.2394	kind	0.0238	-1.2549
anchor	0.0276	-1.441	editor	0.0243	-1.726	journalist	0.0263	-2.729	reporter	0.0298	-2.3259	selectively	0.0135	-1.5338
phoenix	0.0265	-1.442	journalist	0.0198	-2.563	editor	0.0185	-1.661	editor	0.0295	-1.8878	culture	0.0134	-1.3466
producer	0.0257	-1.441	co	0.0117	-0.473	co	0.0171	-0.581	co	0.0288	-0.9164	photog	0.0122	-1.1693
journalist	0.0240	-1.835	host	0.0107	-1.794	breaking	0.0162	-0.596	http	0.0203	-0.3014	married	0.0112	-1.2773
arizona	0.0239	-1.429	pm	0.0098	-1.899	http	0.0146	-0.617	com	0.0195	-1.4846	amazing	0.0107	-0.7359
co	0.0174	-0.758	official	0.0096	-1.602	host	0.0116	-1.453	the	0.0136	-0.2927	no1	0.0102	1.3372
fox	0.0168	-1.091	breaking	0.0095	-0.268	radio	0.0113	-1.109	radio	0.0109	-1.2323	cat	0.0102	-1.2122
kfyi	0.0140	-1.129	producer	0.0091	-1.232	my	0.0100	1.011	events	0.0108	-1.0646	pichona	0.0099	-0.8185
editor	0.0134	-1.100	my	0.0091	0.953	your	0.0100	-0.864	writer	0.0099	-0.0774	news	0.0098	-1.4058
http	0.0124	-0.253	anchor	0.0086	-1.609	com	0.0096	-1.512	my	0.0091	0.7435	page	0.0088	-1.0033
valley	0.0118	-1.267	discuss	0.0085	-0.451	community	0.0091	-1.659	media	0.0086	-0.7891	mad	0.0087	-1.0258
com	0.0117	-0.510	election	0.0079	-1.435	public	0.0083	-1.063	politics	0.0086	0.5008	football	0.0086	-1.2879
nbc	0.0113	-0.725	video	0.0075	-0.781	podcast	0.0082	-1.414	shesourpresident	0.0085	-0.7041	cowboys	0.0085	-1.4590
official	0.0109	-0.905	our	0.0072	-1.061	love	0.0082	1.605	member	0.0085	-1.1687	weakness	0.0083	-0.6781
manager	0.0103	-1.301	http	0.0070	-1.141	official	0.0081	-1.114	host	0.0085	-0.8750	perfect	0.0081	-0.7441
endorsements	0.0096	-0.776	https	0.0070	-0.229	daily	0.0077	-0.950	author	0.0076	-0.2908	gay	0.0078	-0.6285
gilbert	0.0087	-0.872	nycstartup	0.0068	-0.451	journalism	0.0077	-1.334	shopping	0.0075	-0.7425	student	0.0074	-1.1376
breaking	0.0081	-0.646	love	0.0063	0.428	provides	0.0071	-1.195	producer	0.0075	-0.4248	va	0.0070	-1.0441
community	0.0081	-1.099	subjects	0.0063	-0.451	we	0.0069	-0.788	sun	0.0070	-1.0028	friend	0.0067	-0.8256
college	0.0080	-0.892	breakingnews	0.0061	-1.392	https	0.0069	0.073	journalism	0.0068	-1.3141	osu	0.0065	-0.3626
azcentral	0.0079	-0.471	providing	0.0060	-1.282	local	0.0069	-0.849	covering	0.0065	-0.6608	bringing	0.0064	-1.0523
crime	0.0077	-0.600	we	0.0060	-0.630	tv	0.0063	-1.365	historian	0.0060	-0.5329	line	0.0063	-0.5225
retweets	0.0071	-0.512	writer	0.0057	0.010	you	0.0058	0.107	blacklivesmatter	0.0059	-0.6494	project	0.0062	-1.1946
sports	0.0068	-0.523	vonvo	0.0056	-0.451	anchor	0.0056	-1.013	all	0.0058	-0.5009	stories	0.0062	-1.1565
tv	0.0067	-0.870	coverage	0.0055	-1.305	updates	0.0054	-0.700	imstillwithher	0.0058	-0.9763	personality	0.0059	-0.5563
writer	0.0064	-1.165	service	0.0054	-1.277	conservative	0.0054	0.851	tv	0.0058	-0.4538	getting	0.0059	-0.8250
crisis	0.0063	-0.305	latest	0.0054	-0.729	today	0.0052	-1.049	you	0.0057	0.9582	police	0.0058	-1.1148
stories	0.0060	-0.785	radio	0.0053	-0.818	new	0.0052	-0.486	things	0.0056	0.7312	professional	0.0057	-1.0006
network	0.0059	-0.515	new	0.0053	-0.880	based	0.0051	-1.093	freelance	0.0054	-0.2760	happiness	0.0055	-0.9911
weather	0.0059	-0.392	station	0.0051	-0.958	info	0.0050	-1.183	community	0.0054	-1.0805	family	0.0054	-0.4283
guy	0.0056	-0.746	home	0.0050	-1.036	show	0.0050	-0.895	theresistance	0.0053	0.1875	dude	0.0053	-1.3578
truck	0.0056	-1.080	no	0.0049	-1.241	top	0.0050	-0.650	california	0.0052	-0.5120	videogames	0.0053	-0.9686
republic	0.0055	-0.375	channel	0.0049	-1.044	trump	0.0048	0.710	social	0.0052	-0.7697	gonna	0.0053	0.3982
legislative	0.0053	-0.767	organizations	0.0047	-0.684	covering	0.0048	-0.999	former	0.0051	-0.1061	profile	0.0051	-0.7583
mesa	0.0053	-1.066	chief	0.0047	-1.524	writer	0.0048	-0.433	political	0.0050	-0.0833	day	0.0051	-0.7272
asu	0.0052	-0.411	you	0.0046	0.505	usa	0.0047	-1.666	life	0.0049	0.3085	land	0.0051	-0.9104
nd	0.0052	-0.646	lates	0.0046	-0.963	life	0.0046	0.616	public	0.0049	-0.1460	member	0.0049	1.2525
new	0.0051	-0.651	england	0.0046	-0.894	music	0.0045	0.631	me	0.0049	0.3022	resist	0.0049	-0.2272
digital	0.0051	-0.270	covering	0.0046	-1.230	text	0.0045	-0.584	music	0.0048	0.8500	noamnesty	0.0049	-0.0338
twitter	0.0050	-0.229	management	0.0046	-1.168	our	0.0045	-0.519	cool	0.0048	-0.5610	geopolitical	0.0048	-0.8030
my	0.0050	0.566	broadcast	0.0043	-0.830	us	0.0044	-0.549	new	0.0047	-0.5569	looking	0.0047	-1.0887
kgun	0.0048	-0.563	online	0.0043	-0.724	producer	0.0042	-1.063	books	0.0046	-0.4110	today	0.0047	-1.0956
president	0.0048	-0.873	director	0.0042	-0.480	associate	0.0040	-0.997	people	0.0046	0.0127	idk	0.0046	-0.1621
safety	0.0047	-0.355	matter	0.0042	-0.351	digital	0.0039	-0.893	gov	0.0045	-0.5778	company	0.0046	-1.0218
emergency	0.0047	-1.037	talk	0.0042	-0.857	me	0.0039	1.140	love	0.0045	0.6705	manone	0.0046	0.5623
phx	0.0045	-0.556	events	0.0041	-0.346	social	0.0038	-0.467	artist	0.0044	0.6004	pro	0.0046	-0.8893
rts	0.0045	-0.172	former	0.0041	-1.076	tweets	0.0038	-0.841	owner	0.0044	-0.6282	lot	0.0045	-0.4400
call	0.0044	-1.228	tv	0.0040	-0.331	talk	0.0038	-0.926	student	0.0044	1.1484	theresistance	0.0045	-0.5106

Figure 6. Top 50 features with highest coefficients.

know if a feature is indicative of a positive or negative classification, the model must see the feature at least once—ideally multiple times—in the training set. A way to overcome this problem would be to include more negative valued descriptions in the model-training stage and allow it to develop a more expansive negative feature set.

A closer look at the data suggests three causes for misclassification. First, as mentioned above, the restrictive negative feature set results in frequent instances of errors. For example, a profile of an

Table 3. Comparison of predicted truth classification in testing set.

Dataset	True Positive (TP) Prediction	True Negative (TN) Prediction	No. of Positive Samples Model Missed (False Negative)	No, of Negative Samples Model Missed (False Positive)
Mesa	140	19	9	5
Boston	229	152	39	32
Brussels	252	79	35	36
Quebec	311	29	39	18
Random	301	14	46	63

institutional user, who has a true negative value, from the Mesa dataset reads: "*3rd Battalion Fire Radio for Monroe County New York. All information posted is heard from the SCANNER. Nothing is confirmed nor 100% accurate.… #kcco*". To a human coder, this description is obviously used for an organization. The text mentions that the Twitter feed will post fire information "heard from the scanner" and is likely for the residents of Monroe County, New York. Since our model does not reason in such a manner, rather it searches for features that indicate a certain classification, it incorrectly assumes that this is a personal account rather than a Twitter account for an organization. The model was unable to locate features for the negative class because the organizational/institutional users in our dataset predominantly centered around the news media.

Second, there were ambiguous features that could imply a positive classification in one context and a negative classification in another. Since an RF model lacks the ability to reason about features in semantic contexts, ambiguous features are bound to be a source of misclassification. For example, take the personal user from the Boston dataset: *"I love Fox News & Rock 'n Roll and volunteer at animal shelter - rescue is hard but so rewarding too"*. The model falsely made a negative classification with this user. This error is because the terms "Fox" and "News" generally imply a negative classification (or institutional user type). We see in this context, however, that "Fox News" as a feature implies a personal preference. Human coders, unlike machines, could easily pick up on such semantic cues and derive the correct classification.

Third, samples with insufficient information caused misclassification. If a sample lacked enough information to derive a classification with certainty, human coders and machine alike were more prone to misclassify it. For example, the user description with two terms, "*Digital marketing*", from the Boston dataset received a false, negative classification: This is an example of a Twitter account with insufficient information because there is no context for this text. Only a manual checking of the actual profile page could validate the true classification. Such text was problematic because the present features seemed to weakly correlate with a non-personal user type. At first glance, it may seem to a human coder that this is a personal account because it lacks any reference to a specific organization. However, as observed in many examples in our training sets, the features "digital" and "marketing" were generally tied to organizations. For example, a user may say they do "marketing for company X". Hence, to the machine, this could be a good evidence for a negative class since it has, in training, seen these features associated with negative samples. Thus, lack of context was a major source of errors.

Comparisons between two sample groups

Despite some errors, the models overall performed reliably and thus, were used to classify the rest of the unlabeled samples. To address whether the classified two sample groups show systematic differences according to the proposed hypotheses, we ran descriptive post-hoc tests (Chi-square and Kruskal-Wallis H Test) that compared tweet activities between the two sample groups. Figure 7 provides visual summaries of the differences between the two user groups in the four news-related datasets and the Random dataset.

First, an examination of the human-coded data showed mean differences in terms of retweeting across all four news-related datasets. Specifically, the general public user samples were characterized with *more* retweeting than institutional user samples (H1). Also, Boston, Brussels, and Quebec datasets showed that general users tweeted URLs *less frequently* when referring external information than institutional users (H2). Although the Mesa dataset was not statistically significant for H2, the pattern found in the tweets was consistent with the other three datasets. In terms of language use, general public users showed significantly *less* analytic tones than institutional users across Mesa, Boston, and Brussels datasets (H4). Although the Quebec dataset was not statistically significant regarding analytic languages, the pattern found in the tweets related to the event was consistent with the other three datasets. The use of affective language was significantly *higher* in the general user tweets than institutional user tweets only in the Boston dataset (H5). While the use of affective language was not significant for other datasets, the pattern was, again, consistent across the datasets.

Second, an examination of the machine-labeled samples reaffirmed the differences found from the hand-coded samples in terms of H1, H2, H4, and H5. That is, the variables that were not significant with the hand-coded dataset showed significant differences between the two samples in the machine-labeled datasets. For example, use of affective language became significant in Mesa, Boston, and Quebec datasets when analyzed in the machine-labeled samples, with consistent mean difference patterns. The statistical differences in language styles for general users—lower in analytic tones and higher in affective tones than institutional users—were resonant with previous anecdotal discussions that general users showed more affect in their own content on Twitter than media users

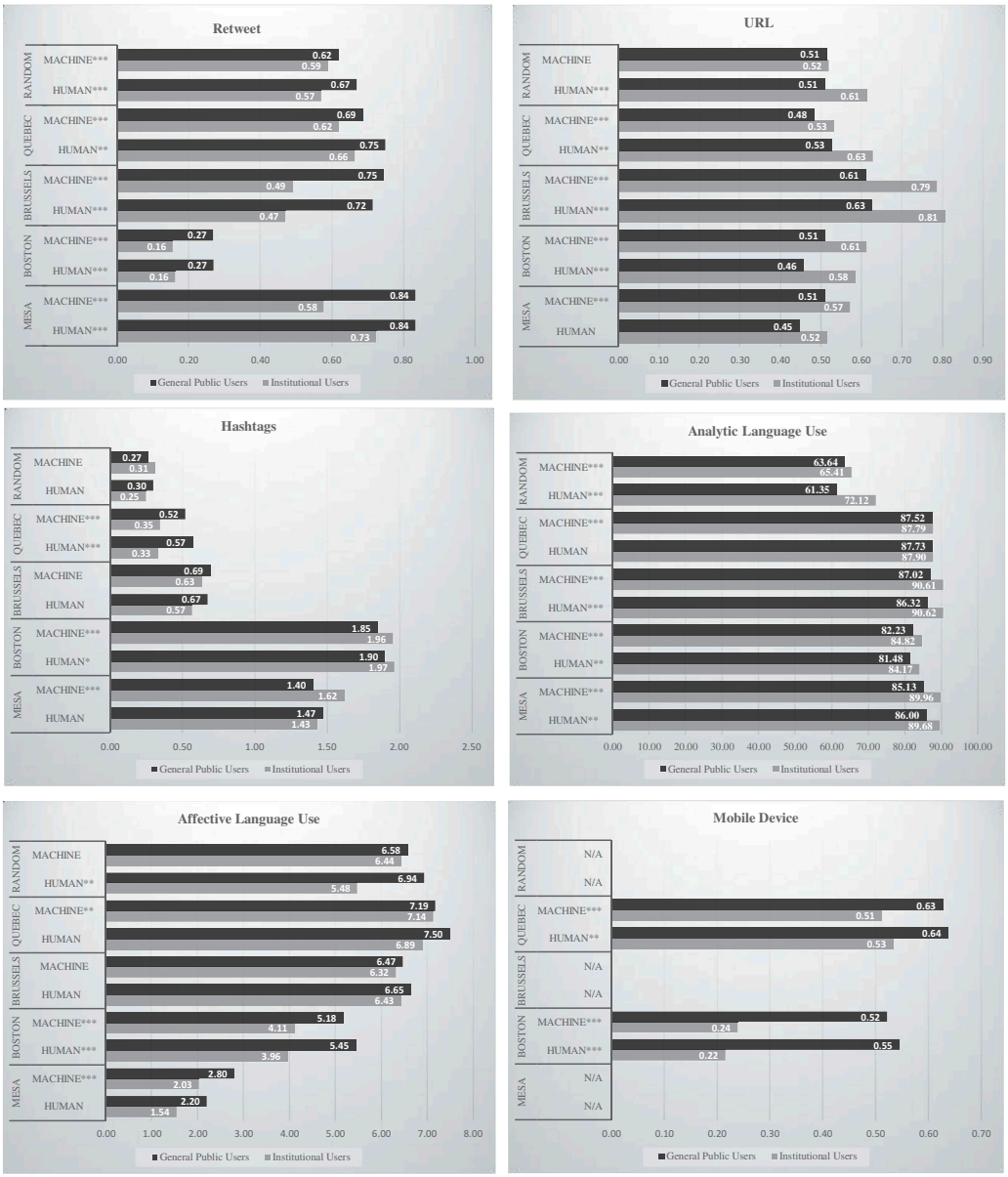


Figure 7. Summary of mean differences between general public users and institutional users. *** $p < .001$; ** $p < .01$; * $p < .05$; Retweet, URL, Mobile Device were tested by Pearson $X^2(1)$. The rest were tested by Kruskal-Wallis rank $X^2(1)$ (with ties considered).

(Papacharissi, 2015). Our findings not only reiterate these previous discussions but also emphasize the importance of separating samples into general users and institutional users to truly understand who is tweeting what and how.

Meanwhile, hash-tagging—as hypothesized as H3—became significantly different in the machine-classified samples. However, the directionality of the mean differences were *inconsistent* across different datasets, thus preventing us from reporting systematic differences between general public and institutional users regarding the use of hashtags. Prior scholarship may explain such inconclusive findings regarding hasthtags: they have shown that news media were not "skillful" in using hashtags and keywords for their tweets even if some news outlets tweeted rather frequently (Engesser & Humprecht, 2015, p. 514). Their discussion implies that institutional users—notably media professionals—are not distinctive from general public users due to the lack of strategic use of hashtags.

Next, we tested H6—difference in tweeting patterns of general public users and institutional users based on the device used to tweet—for the Boston and Quebec events. The results suggested that general public users used a mobile device *twice as often* as institutional users to tweet their posts (H6). This result implies the possibility that "the type of people who tweet from mobile devices are qualitatively *different* from those who tweet from web-based platforms" (Murthy et al., 2015, p. 834, *italics* added). The difference in device usage was also significant in the Quebec dataset, reiterating the results from the Boston dataset, although the gap between the two groups was reduced to 10%.

Finally, we replicated the post-hoc tests on the Random dataset to address the generalizability of these findings. The results from Random dataset remained largely the same as the findings from the four violent-event based datasets. That is, general users *retweeted more; used URLs less; and used less analytic language yet more affective language* than institutional users. While some of the variables were significant only for either the machine-labeled dataset or the human-coded dataset, the patterns of mean differences were persistent. There was no difference in hashtagging between general users and institutional users in the Random dataset, that also reaffirmed the inconsistent results for this variable as found in the other four datasets.

In summary, the overall tweeting patterns revealed systematic differences between the general public user samples and institutional user samples in terms of retweets, use of URLs, and language use, but not in terms of hashtag use.

Discussion and conclusion

This article intends to address sampling issues faced by communication scholars engaged in social media research. Data gathered from Twitter—or any other social media platform—usually involves an examination of complex sets of information. As highlighted in the literature section, the heterogeneous nature of data could induce various sampling biases, specifically predefined sampling frames, engineered noises and proxy misspecifications. Whereas many issues related to these biases are unfortunately beyond the researchers' control, this study suggests that some shortcomings may be improved upon with a computational approach. Particularly, this study highlighted the "proxy-population mismatch" issue, and demonstrated the use of the Random Forest classifier to improve the proxy-population match by disentangling general public users from institutional users in a general Twitter sample.

We manually prepared the training and testing sets comprising 7,000 human-coded Twitter profiles, and then used them to train the machine-learning model to computationally distinguish between general and institutional users. The preparation of coded data followed a rule-based approach, that was later validated by an additional review of the actual user profile page where available. We applied the trained models to four different violence-news related Twitter datasets that centered around a local event (Mesa), a national event (Boston) and two international events (Brussels and Quebec). To address the generalizability of these results beyond specific news contexts,

we added a Random dataset and replicated the whole modeling and analysis processes with randomly selected tweets on uneventful days.

The results are evidence that RF modeling may be a useful approach to distinguish between the two types of Twitter account holders within a given news event data, and possibly across news events of a similar nature. The model performed well across all four news datasets: within the Mesa shooting dataset, the performance was above 90% for accuracy, precision, recall, and F1 scores; the results for the other three were also solid, with a performance of higher than 80% for accuracy, precision, recall, and F1. Moreover, cross-validation of the model across the four datasets also showed satisfactory results, suggesting the possibility of using the machine-learning model trained with these datasets for future research on Twitter datasets that pertain to other violence-related news events. The Boston dataset, however, showed results with lower accuracy than others when tested for cross-validation. This result could be due to a particularly large portion of the negative class (i.e., institutional users) contained in the Boston dataset—almost 40%—compared to the other three datasets.

The high performance of our model could be good news especially for social media researcher who would like to implement computational process for a better large-scale sampling. We intended to develop a machine learning model that would be useful in the sampling stage of their research. Thus, one of our practical goals was to reduce the modeling cost by minimizing data-related preprocessing steps and using simple algorithms to do so. While more advanced feature engineering and complex algorithms could have resulted in higher performances, the simplicity of the TF-based feature sets and use of the RF classifier reduced the time and effort necessary to classify the sample in different categories.

Nonetheless, it is important to discuss the limitations of our model. First, this study primarily focused on the proxy-population mismatch issue, using "front-end" data composed of publicly available tweets and user profile data. It could not address problems inherent in a predefined sampling frame or engineered noise issues. Specifically, the most accurate information about Twitter user demographics could reside in proprietary "back-end" data. Restricted access to this back-end data is one intractable limitations that social media researchers often encounter when they estimate the population of interest on a platform like Twitter. Additionally, StreamAPI-based data collection, as opposed to purchasing the data via Firehose, compromises transparency regarding the algorithms Twitter uses to predefine the sample frame available via StreamAPI. Such limitations are a result of how Twitter shares its data with academics and due to their proprietary nature, beyond the scope of this study.

Another limitation pertains with the labeling approach we took when preparing train and test datasets. That is, the profile text-based classification could miss subtle, contextual cues that differentiate between institutional and personal users. For example, the textual description in a user's profile may be insufficient to conclude whether the user belongs to the general public or institutional category. To overcome this stumbling block, future studies may want to adopt the manual checkup of the actual Twitter profiles—which we did for validation purposes—as the ground truth label, and add some additional non-textual features (e.g., profile photo, etc.) when building feature sets. Also, researchers interested in this research might consider a higher n-gram embedding of the data. This procedure could, however, lead to substantially greater computational complexity and possibly a model that over fits the data. A more suitable approach would be to build a new model that considers features within their context more naturally. An example of such a model would be a word2Vec neural network, one that learns vector representations of phrases rather than each term (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The current study did not try this model, which is worth further investigation.

Our error analysis suggests that a small size of the negative training set (=institutional users) resulted in a narrow range of features from which the model could learn only partially about institutional users. This led to a high chance of False Positive classification cases. Although our feature set was a relatively solid representation of media/journalism-related users, it could have missed other non-media institutional users. This could be a reason why institutional users were consistently underrepresented across all machine-labeled datasets: Institutional users were less prevalent in all machine-labeled datasets than their

counterpart hand-coded datasets. One practical solution to this issue would be to use a training set that has a more diverse representation of institutional users. One might incorporate negative labels from Random datasets into the training set for more accuracy in this regard.

Overall, our study suggests the possibility of a nontrivial portion of institutional user tweets included in a sampling frame if they are not carefully filtered out. Even in the Random data, we observed over 17% of messages were identified as institutional tweets. Filtered sampling could be particularly advantageous in some research contexts where general public user behaviors are of particular interest. Specifically, general public users are likely to (i) retweet others' messages more, (ii) hyperlink external information less, (iii) use lesser analytic and more affective language, and (iv) tweet more from mobile devices than institutional users. These findings imply that researchers' failure to filter out non-general users from a Twitter sample would have confounding effects on a study's results, especially when the purpose is to understand general users' behaviors.

Our finding of general users using the singular pronoun "my/me" more than institutional users adds nuance to Murthy et al. (2015) finding of tweets from mobile devices incorporating more egocentric language than tweets from web devices. A more critical question, though, will be whether the use of egocentric language is a result of the affordance of mobile devices or the nature of the population that primarily tweets from mobile devices. On a related note, only about 20% of non-general/institutional users had tweeted from mobile devices in 2013, as seen in the Boston dataset. The number of cases related to mobile use when tweeting, however, increased more than twice, up to about 50%, during the Quebec mosque attack in 2017. This drastic change within the institutional user samples is interesting, and worthy of further investigation. Researchers could examine whether such change influences the ways in which institutional users frame their messages.

In summary, our study shows that RF modeling offers an effective way of "slicing" a large social media dataset. The RF classifier is useful when separating binary classes under the linearly non-separable data condition. With great promise, however, comes great expectation and the method shown in this article has its limitations. We tested RF modeling for a binary classification using only the Twitter profile texts. Future studies may extend this line of research by (i) investigating multi-label classification problems, and (ii) by addressing more complex features such as network structures or other information available in user accounts, as mentioned earlier. Another problem that the social media research community should ponder is identifying and excluding bot activities when studying human behaviors. Bots are an entrenched problem that not only creates engineered noises but also aggravates the proxy-population mismatch. While it is beyond the scope of our project to disambiguate the human-like bot profiles, there is an urgent need for future research to develop advanced and sophisticated methods in this regard. With recommended future research alternatives, scholars could further enhance the utility of advanced computational techniques in communication research.

Notes

1. Recall, Precision, and Accuracy are the standard metrics to validate ML results. There are four possible classification outcomes: (1) True Positive (TP); (2) False Negative (FN); (3) False Positive (FP); and (4) True Negative (TN). Recall is the rate of correctly labeling general publics out of all the instances supposed to belong to general publics, computed as TP/(TP + FN). Precision is the rate of including correctly labeled general public users from all instances labeled as general publics, computed as TP/(TP + FP). Accuracy is the rate of correct labels of both, the general public and institutional users from all labels, computed as (TP+TN)/(TP + FN + FP + TN). F1 score is the harmonic mean of precision and recall, computed as 2 * (Precision * Recall)/(Precision + Recall).
2. More formally, the GI for a classifier with J labels can be computed via the following:

$$\sum_{i=0}^{J} P_i(1 - P_i) = \sum_{i=0}^{J} P_i - P_i^2 ,$$

where $P_i :=$ probability a randomly chosen object is classified with label i.

For the source codes and label data, please see the project repository at https://github.com/jpriniski/TwitterClassification

Funding

This work was supported by the 2017 Emerging Scholars Award from Association for Education in Journalism and Mass Communication.

ORCID

K. Hazel Kwon http://orcid.org/0000-0001-7414-6959

References

Abokhodair, N., Yoo, D., & McDonald, D. W. (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* (pp. 839–851). New York, NY: ACM. doi:https://doi.org/10.1145/2675133.2675208

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Breiman, L. (1984). *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC.

Breiman, L. (2001). *Random Forests*. Retrieved April 10, 2017, from https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

Brooks, B., Hogan, B., Ellison, N., Lampe, C., & Vitak, J. (2014). Assessing structural correlates to social capital in Facebook ego networks. *Social Networks*, *38*, 1–15.

Burrows, R., & Savage, M. (2014). After the crisis? Big data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1). doi:10.1177/2053951714540280

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

Cohen, R., & Ruth, D. (2013). Classifying political orientation on Twitter: It's not easy! In Seventh International AAAI Conference on Weblogs and Social Media (pp. 91–99). Cambridge, MA: AAAI Press.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013, June). Carmen: A twitter geolocation system with applications to public health. *Papers from the AAAI workshop on expanding the boundaries of health informatics using* (pp. 20–24). Bellevue, WA, USA. Association for the Advancement of Artificial (AAA).

Driscoll, K., & Walker, S. (2014). Big data, big questions, working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, *8*, 1745–1764.

Emery, S. L., Szczypka, G., Abril, E. P., Kim, Y., & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in Tweets about the Tips campaign. *Journal of Communication*, *64*(2), 278–295.

Engesser, S., & Humprecht, E. (2015). Frequency or skillfulness. *Journalism Studies*, *16*(4), 513–529.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16–27.

Hargittai, E. (2015). Is bigger always better? potential biases of Big data derived from Social Network Sites. *The Annals of the American Academy of Political and Social Science*, *659*(1), 63–76.

Hargittai, E., & Litt, E. (2011). The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media & Society*, *13*(5), 824–842.

Hermida, A. (2010). Twittering the news. *Journalism Practice*, *4*(3), 297–308.

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather wweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 40–60.

Jackson, S. J., & Foucault Welles, B. (2015). Hijacking #myNYPD: Social media dissent and networked counterpublics. *Journal of Communication*, *65*(6), 932–952.

Kwon, K. H., Chadha, M., & Pellizzaro, K. (2017). Proximity and terrorism news in social media: A construal-level theoretical approach to networked framing of terrorism in Twitter. *Mass Communication and Society*, *20*(6), 869–894.

Kwon, K. H., Stefanone, M. A., & Barnett, G. A. (2014). Social network influence on online behavioral choices: Exploring group formation on Social Network Sites. *American Behavioral Scientist*, *58*(10), 1345–1360.

Lee, K., Eoff, B., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on Twitter. In Fifth International AAAI Conference on Weblogs and Social Media. Barcelona, Spain.

Loh, W. (2011). *Classification and Regression Trees*. Retrieved from http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf 10.1002/widm.8

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. arXiv:1306.5204 [Physics]. Retrieved from http://arxiv.org/abs/1306.5204

Murthy, D., Bowman, S., Gross, A. J., & McGarry, M. (2015). Do we tweet differently from our mobile devices? A study of language differences on mobile and web-based Twitter platforms. *Journal of Communication*, *65*(5), 816–837.

Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics*. New York, NY: Oxford University Press.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates. www.LIWC.net

Ratkiewicz, J., Conover, M., Miess, M., Goncalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. In *WWW 2011 Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 249–252). Hyderabad, India: ACM.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Salganik, M. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data & Society*, 1(2), 1–11.

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication, 67*, (2), 233–255. doi:10.1111/jcom.12284

Smith, M., Ceni, A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., & Dunne, C., (2010). NodeXL: A free and open network overview, discovery and exploration add-in for Excel 2007/2010/2013/2016, http://nodexl.codeplex.com/. Social Media Research Foundation

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. arXiv:1111.4503 [Physics]. Retrieved from http://arxiv.org/abs/1111.4503

Vaccari, C., Chadwick, A., & O'Loughlin, B. (2015). Dual screening the political: Media events, social media, and citizen engagement. *Journal of Communication*, *65*(6), 1041–1061.

Vargo, C. J., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network issue agendas on Twitter during the 2012 U.S. presidential election. *Journal of Communication*, *64*(2), 296–316.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. arXiv:1703.03107 [Cs]. Retrieved from http://arxiv.org/abs/1703.03107

Woodford, D., Walker, S., & Paul, A. (2013). Slicing big data - Twitter, gambling and time sensitive information. In *Selected Papers of Internet Research* . 14.0. Denver, CO. Association of Internet Research. Retrieved from http://spir.aoir.org/index.php/spir/article/view/914

Zhou, W.-X., Sornette, D., Hill, R. A., & Dunbar, R. I. M. (2005). Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1561), 439–444.