Social Media for Large Studies of Behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards.

Derek Ruths and Jürgen Pfeffer

On November 3, 1948, the day after Harry Truman won the United States presidential elections, the Chicago Tribune published one of the most famous erroneous headlines in newspaper history: “Dewey Defeats Truman” (1,2). Rather than permanently discrediting the practice of polling, this event led to the development of more sophisticated techniques and higher standards that produce the more accurate and statistically rigorous polls conducted today (*3*).

Now we are poised at a similar technological inflection point. Recent years have seen the rise of online personal and social data for the study of human behavior. Powerful computational resources combined with the availability of massive social media datasets have given rise to a growing body of work that uses a combination of machine learning, natural language processing, network analysis, and statistics for the measurement of population structure and human behavior at unprecedented scale. However, mounting evidence suggests that many of the forecasts and analyses being produced misrepresent the real worl*d (5,6,7*). Here we highlight issues that are endemic to the study of human behavior through large-scale social media datasets and discuss strategies that can be used to address them. While some of the issues raised are very basic (and long-studied) in the social sciences, the new kinds of data and the entry of a variety of communities of researchers into the field make these issues worth revisiting and updating.

Representation of Human Populations

Population bias. A common, assumption underlying many large-scale social media-based studies of human behavior is that a large-enough sample of users will drown out noise introduced by peculiarities of the platform’s population (*8*). However, substantial population biases exist that vary across different social media platforms (*9*). For instance, Instagram is “especially appealing to adults aged 18-29, African-American, Latinos, women, urban residents” (10) whereas Pinterest is dominated by females, age between 25 - 34, with an average annual household income of $100,000 (*11*). Despite these sampling biases being built into platforms used for scientific studies, they are rarely corrected for (if even acknowledged).

Proprietary algorithms for public data. The dependence of most social media research on public data feeds raises serious platform-specific sampling problems. For example, recent work shows evidence that the highest volume source of public Twitter data, which is used by thousands of researchers worldwide, is not an accurate representation of the overall platform’s data (*12*). Furthermore, researchers are left in the dark about when and how social media providers change the sampling/filtering of their data streams. So long as the algorithms and processes that govern these public data releases are largely dynamic, proprietary, and secret or undocumented, designing reliable and reproducible studies of human behavior that correctly account for the resulting biases will be difficult, if not impossible. Academic efforts to characterize aspects of the behavior of such proprietary systems can provide details needed to begin reporting biases.

The rise of “embedded researchers” (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms and resources) is creating a divided social media research community. Such researchers, for example, can see a platform’s inner workings and make accommodations, but may not be able to reveal their corrections or the data used to generate their findings.

Representation of Human Behavior

Human behavior and online platform design. Many social forces that drive the formation and dynamics of human behavior and relations have been intensively studied and are well-known (*13,14,15*). For instance, homophily (“birds of a feather flock together”), transitivity (“the friend of a friend is a friend”), and propinquity (“those close by, form a tie”) are all known by designers of social media platforms and, to increase platform use and adoption, have been incorporated in their link suggestion algorithms. Thus, it may be necessary to untangle psycho-social from platform-driven behavior. Unfortunately, few studies attempt this.

Social platforms also implicitly target and capture human behavior according to behavioral norms that develop around and as a result of the specific platforms. For instance, the ways in which users view Twitter as a space for political discourse affects how representative political content will be. The challenge of accounting for platform-specific behavioral norms is compounded by their temporal nature: they change with shifts in population composition, the rise and fall of other platforms, and current events (e.g., revelations concerning interest and tracking of social media platforms by intelligence services). New methodologies are needed to disentangle the human from the platform. In the absence of such methods, we must rely on assessments of where such entanglements likely occur.

Distortion of human behavior. Developers of online social platforms are building tools to serve a specific, practical purpose - not necessarily to represent social behavior or provide good data for research. This raises the question of whether the way data are stored and served destroys aspects of the human behavior of interest. For instance, Google stores and reports final searches submitted, *after* auto-completion is done, as opposed to the text actually typed by the user (*6*); Twitter dismantles retweet chains by connecting every retweet back to the original source (rather than the post that triggered that retweet). There are valid, practical reasons for platforms to make such design decisions, but in many cases these either obscure or lose important aspects of the underlying human behavior. Quantifying and, if possible, correcting for these storage and access policies should be part of the dataset reporting and curation process.

Non-humans in large-scale studies. Despite attempts by platform designers to police accounts, there are large populations of spammers and bots masquerading as "normal" humans on all major online social platforms. Moreover, many prominent individuals maintain social media accounts that are professionally managed to create a constructed image or even behave so as to strategically influence other users. It is currently impossible to accurately remove or correct for the vast majority of such distortions.

Issues with Methods

Proxy population mismatch. Every 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 to be studied are commonly instead, for example the set of all Facebook users who report attending a UC school. However, the quantitative relationship between the proxy and original populations studied, typically, is unknown, making this practice a source of potentially serious bias. A recent study, for example, revealed that this proxy effect has caused substantially incorrect estimates of political orientation on Twitter (*7*).

Incomparability of methods and data. With few exceptions, the terms of usage for social media platforms forbid the retention or sharing of datasets collected from their sites. As a result, canonical datasets for the evaluation and comparison of computational and statistical methods - common in many other fields - largely do not exist in the social media community. Furthermore, few researchers publish code implementing their methods. The result is a culture in which new methods are introduced (and often touted as being "better") without having been directly compared to existing methods on a single dataset. Given platforms' understandable sensitivity to user privacy and the competitive value of their data, the research community will likely improve method and result comparison issues more quickly by focusing on enforcing the sharing of methods at publication time.

Multiple comparison problems. The body of social media analysis that concerns the development of user/content classification and prediction has unaddressed issues with overfitting. Specifically, when building a computational machine that recognizes two or more classes (of users, for example), it is customary to introduce tens to hundreds of features as the basis for the classifier. At the very least, the performance of the classifier should take into account the number of features being used. Of greater concern, however, is the extent to which the classifier performance is a result of "feature hunting" - testing feature after feature until one is found which delivers significant performance on the specific dataset. Standard practices from machine learning of reporting the p-value for classifiers based on the number of features involved as well as keeping a dataset independent of the training set for final classifier evaluation would represent significant strides towards addressing these issues (*16*).

Multiple hypothesis testing. In an academic culture that celebrates only positive findings, a meta-issue emerges as multiple groups report successes in modeling or predicting a specific social phenomenon. If we do not observe the failed studies, then we cannot assess the extent to which successful findings are the result of random chance. This issue has been observed already in the literature on predicting political election outcomes with Twitter (*4*). We are not the only field struggling with this issue (*17*). Solutions to this problem could involve enabling the publication of negative results or requiring the use of more datasets in a single study (so as to permit the calculation of a significance score within the study itself).

Conclusions

The biases and issues highlighted above will not affect all research in the same way. Well-reasoned judgment on the part of authors, reviewers, and editors are warranted here. Many of the issues discussed have well-known solutions contributed by other fields such as epidemiology, statistics, and machine learning. In some cases, the solutions are difficult to fit with practical realities (e.g., as in the case of proper significance testing) while in other cases the community simply has not broadly adopted best practices (e.g., independent datasets for testing machine learning techniques) or the existing solutions may be subject to biases of their own. Regardless, a crucial step is to resolve the disconnect that exists between this research community and other (often related) fields with methods and practices for managing analytical bias.

Moreover, while the issues highlighted above all have different origins and specific solutions, they share in common the need for increased awareness of what is actually being analyzed when working with social media data.

Table 1: Checklist of approaches for large-scale social media studies of human behavior (Further discussion in SOM)

Data Collection

- ❏ 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- ❏ 2. Quantifies biases of available data (access constraints, platform-side filtering)
- ❏ 3. Quantifies proxy population biases/mismatches

Methods

- ❏ 4. Applies filters/corrects for non-human accounts in data
- ❏ 5. Accounts for platform and proxy population biases
  - ❏ a. Corrects for platform-specific and proxy population biases

    OR
  - ❏ b. Tests robustness of findings
- ❏ 6. Accounts for platform-specific algorithms
  - ❏ a. Shows results for more than one platform

    OR
  - ❏ b. Shows results for time-separated datasets from the same platform
- ❏ 7. For new methods: compares results to existing methods on the same data
- ❏ 8. For new social phenomena or methods or classifiers: reports performance on two or more distinct datasets (one of which was not used during classifier development or design)

REFERENCES

1. This was not the first or last such erroneous prediction. E.g., the Literary Digest on the 1936 US Presidential election.

2. F. Mosteller, H. Hyman, P. J. McCarthy, E. S. Marks, and D. B. Truman (1949). The Pre-Election Polls of 1948. Bulletin 60, Social Science Research Council, New York.

3. I. Crespi (1989). Public opinion, polls, and democracy. Boulder, CO. Westview Press.

4. H. Schoen, and D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, P. A. Gloor (2013). The power of prediction with social media. Internet Research 23(5):528-543.

5. Z. Tufekci. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In ICWSM ’14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.

6. D. Lazer, R. Kennedy, G. King, A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science 343:1203.

7. R. Cohen and D. Ruths (2013). Political Orientation Inference on Twitter: It’s not easy! Proceedings of ICWSM ’13 (pp. 91-99).

8. V. Mayer-Schoenberger and K. Cukier (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. Houghton Mifflin Harcourt.

9. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist (2011). Understanding the demographics of Twitter users. ICWSM ’11 (pp. 554-557).

10. http://www.pewinternet.org/Reports/2013 /Social-media-users.aspx

11. http://www.pinterest.com/pin/ 234257618087475827/

12.F. Morstatter, J. Pfeffer, and H. Liu (2014). When is it Biased? Assessing the Representativeness of Twitter's Streaming API. WWW 2014, Web Science Track Pages 555-556).

13.M. McPherson, L. Smith-Lovin, and J. M. Cook (2001). Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27:415-444.

14.F. Heider (1946). Attitudes and Cognitive Organizations, *Journal of Psychology* **21**, 107-112.

15.L. Festinger, S. Schachter, K. Back (1950). The Spatial Ecology of Group Formation. In L. Festinger, S. Schachter, & K. Back (Eds.), *Social Pressure in Informal Groups* (p. Chapter 4). Cambridge, MA: MIT Press.

16.S. J. Russell and P. Norvig (2003). Artificial Intelligence: A Modern Approach. Pearson Education.

17.J. P. A Ioannidis (2005). Why Most Published Research Findings are False. *PLoS Medicine* **2**, e124. doi:10.1371/journal.pmed.0020124

Supplementary Information Text

Table 1 is provided as a guideline for researchers, authors, reviewers, and editors. At the outset, it is important to emphasize that, given the implicitly creative and dynamic nature of research, there is no easy way to stipulate rules which all work should follow. Nonetheless, the checklist distills the key challenges discussed in the paper into operational pieces that should be present in every paper in some form. And while not all studies may be able to (or need not) decisively handle each item, authors should explain why their study cannot or should not be held to particular standards.

1. Quantifies relevant platform-specific biases. The goal here is to ensure that the authors have anticipated and readers are aware of all possible platform-specific factors that could confound the study including what and how user data is stored by the platform (6), the demographics of the user base (10,11), and user behavior that is particular to the platform. Ideally, such quantification will be accurate, though it is possible that some biases will be impossible to precisely quantify.

2. Quantifies relevant biases of available data. If APIs or web crawlers are used to collect data instead of unfettered access to data maintained by social media platforms, the explicit or implicit data bias that comes with the data collection approach should be discussed (12).

3. Quantifies relevant proxy population biases/mismatches. Most (if not all) social media studies make at least some use of proxy populations (7) - even if the “proxy” is simply a direct attempt to operationalize the definition of a group. As in the previous items, an explicit and exhaustive enumeration of the sources of bias both demonstrates the authors’ awareness of these biases and alerts the reader to their possible presence (making a more critical reading of the work possible).

4. Applies filters/corrects for non-humans in data (where appropriate). When “authentic” human users are needed for a study, a credible attempt to remove bots, prominent personalities (e.g., politicians and celebrities), and organization accounts from the dataset should be used. Since the detection of non-human accounts is an active area of research, an argument should be made that the appropriate state-of-the-art methods have been used.

5. Accounts for platform and proxy population biases. At a very high-level, there are two ways to account for biases. The more preferable approach is to make appropriate statistical corrections that are informed by known biases in the data (item 5.a.). It is possible that such corrections cannot be made, either due to lack of knowledge of the exact amount of biasing or the complexity of methods/analyses being used. Under such circumstances, an alternative is to test the robustness of the findings of the study (item 5.b.) by re-running the analyses on a dataset that explicitly introduces/removes biasing factors. If findings (whether methodological advances or analytical insights) are to be accepted, they must pass such robustness tests.

6. Account for platform-specific algorithms. Studies of human behavior may focus on platform-specific or platform-independent phenomena. Both must contend with the way that platform-specific algorithms may affect results. Those which assert a general phenomenon present in social media (e.g., social media can be used to analyze mood changes) must show that their core conclusions can be supported by results on multiple platforms (item 6.a.). Those which consider platform-specific behavior must address the possibility that the platform has changed its algorithms and site design over time - motivating the need for showing that results are reproducible over time (item 6.b.). It is worth noting that platform-specific findings are intrinsically more limited in terms of their ability to make broad claims about the nature of human behavior - and, thus, generalizing claims emanating from such studies should be carefully considered within the context of the data and methods.

7. Compares results to existing methods on the same data. The only way to definitively show that a method has improved over the state-of-the-art is to show that the method actually outperforms existing methods on the same dataset. Note that such comparisons are only necessary in studies which are introducing new methods.

8. Reports performance of new finding on two or more distinct datasets. This practice simultaneously addresses concerns about overfitting specific datasets and multiple hypothesis testing. Running a trained model on multiple distinct datasets (e.g., collected at different time periods, using different methods, or on different platforms) reduces the risk that performance achieved is the result of fitting the data to a particular dataset or experimental condition. Moreover, given that negative findings are rarely published and are subject to selective bias themselves, evaluation on multiple datasets is a practical way of demonstrating that findings are not the result of “running the right method in the right place at the right time.”

A holdout dataset is an approach (complementary to those discussed above) for ensuring that a classifier of interest has not been overly tuned to work well on the dataset(s) used elsewhere in the study (16). This is a special dataset which was not consulted or used in any way during the development of the classifier (note that this is different from the validation dataset, also called a hold-out dataset, used during cross-fold validation).

Final thoughts

A fair retort that can be made to guidelines like those above is that there is “not enough data” available. Indeed, this can be a very real problem for researchers. However, this objection needs to be very carefully examined and, in many cases, retired. Allowing the publication of work which is based on insufficient data must be weighed against the substantial risk that erroneous findings or results with ambiguous generalizability are entered into the academic record and become the basis for future work. Often this problem can be addressed by scaling back the scope or reframing the central hypothesis of a study to address a more specific aspect of human behavior on a given platform.

Also notable is that a number of the recommendations above require authors to comment on and attempt to correct for sources of bias (in particular, items 1, 2, and 3; 5.a and 5.b, respectively), many of which may not be entirely known or knowable. On the one hand, it is important to recognize that such circumstances do not justify the omission of the discussion of the biases, but rather demand that authors be forthcoming about why such biases cannot be characterized more precisely. This said, it is important that reviewers and editors be careful not to overly penalize authors for pointing out biases for which they cannot entirely account. While it is tempting to view such admissions as weaknesses of the study, bringing such “known unknowns” to light creates new research directions and advances the state of the field; rejecting such work will increasingly create a research culture in which important sources of difficult-to-assess bias are unnecessarily hidden and remain unaddressed. Certainly, moving towards greater transparency will involve the energy and commitment of the entire research community, but will, ultimately, contribute to substantial advances in the study of human behavior on and through social media platforms.