Reliability of Data Collection Methods in Social Media Research

Giorgos Cheliotis
Communications and New Media
National University of Singapore
gcheliotis@gmail.com

Xuesong Lu
Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
xuesong.lu@epfl.ch

Yi Song
School of Computing
National University of Singapore
songyi@nus.edu.sg

Abstract

In this paper we argue for the systematic assessment and reporting of the reliability of data collection methods that rely on the automated collection of data from third party social media sites, where the researcher does not have direct access to meaning or control over data quality. Building on a long tradition of reliability assessment in the sciences and specifically in social science, we propose methods for the assessment of reliability, also for textual data that is increasingly mined from social media for the purposes of studying online populations.

Introduction

The Web and in particular social media have proven to be real treasure troves of information on the behavior of individuals, as well as the collective behaviors that emerge from the interactions of many. Most research that mines the Web for such data relies to various degrees on the accuracy and appropriate use of large datasets that were sourced from a third party's website, either by crawling relevant webpages for information, or by querying public APIs. Although there are benefits to observation of social behavior on the web – especially if we consider that it is difficult to produce proper random samples of online populations (Fricker and Schonlau 2012; Wright 2005) – it is dangerous to assume without proof (Borgatti et al. 2009) that the data we collect from the Web and from social media in particular is free of the measurement error that also plagues other methods of data collection, e.g. survey instruments commonly employed in the social sciences.

This also holds true when data is collected using own software scripts tailored for this purpose, or third party tools that aid the researcher in automating the data collection process. Our tools may be well specified, but their output will be stochastic, not deterministic, thus subject to random error, or, if we are not careful, also systematic error. Recent research on this topic has focused on questions of sampling bias and on the validity of measures computed from social media datasets. We are concerned here with the equally important but oft neglected issue of reliability. We ask: "even when we may be able to account for sampling bias and the measures we produce are theoretically valid, have we checked to see if our data collection tools are even reliable, in that they yield consistent datasets?"

Literature Review

Issues of random and systematic measurement error are addressed in the social sciences using multiple indicators, to reduce uncertainty in the data collection process and mitigate errors that may eventually lead to incorrect conclusions about the populations under study (Curtis and Elton 1962; Tufekci 2014). It is for example common to use multiple indicators in survey research (Rea and Parker 2005), as it is generally prone to a number of random errors and bias, and there is no direct access to meaning: the researcher depends on what the survey respondent will reveal, which in turn depends on many factors, including the reliability of the measurement instrument. But direct access to meaning is also not guaranteed for software and database systems. The instructions of a computer program can be defined in certain terms, but the communication between different systems on the path from an online database on a server, to the client executing the data collection program, as well as the interpretation of the results by the human analyst, are subject to a number of errors that in many ways reflect the types of errors inherent in research that relies on the active participation of human subjects.

In test theory, reliability concerns the extent to which different measurements are in agreement with one another (Carmines and Zeller 1979). Correlation estimates are the foundation of any reliability measurement. Related to reliability is the concept of validity. While reliability is about the accuracy and replicability of the measurement, validity is about whether the data collected can yield a valid meas-

ure of a theoretical construct that is pertinent to our research question. A measurement instrument needs to be *both reliable and valid*. In this paper we focus our attention on the assessment of reliability. A related, but distinct issue is the choice of sampling method and the mitigation of sampling bias.

Although largely ignored in many published studies, as the field matures, there is growing awareness of these issues in social media research. Tufekci (2014) identifies a number of common issues in such research: the model organism problem; selecting on dependent variables; the denominator problem in sampling; and the prevalence of single platform studies. She questions the reliability of data collection methods utilizing automated scripts to query public APIs and among other suggestions, she recommends comparing the results of multiple data collection methods. González-Bailón et al. (2012) compare two datasets obtained by querying the Twitter Stream and Search APIs respectively, with an aim towards comparing the coverage of the two data collection methods and identifying sampling bias. The authors find that the Search API yields a smaller dataset than the Stream API, which also results in over- or under-estimation of various network metrics they compute based on networks constructed of tweet mentions and retweets.

In another effort (Morstatter et al. 2013), the authors compare a sample retrieved using the Twitter Stream API with results obtained directly from the Twitter *firehose*, a means of direct access to Twitter's data that is only available at a premium and thus not commonly employed in Internet research. They find that the Streaming API's coverage varies on a daily basis, and that in some cases it performs worse than comparable datasets randomly sampled from the firehose.

Problem statement

We complement the above literature, which has mostly dealt with the issues of sampling bias and validity, by focusing on the issue of reliability. As already discussed, these are related, but distinct issues. In order to produce scientifically rigorous research using data mined from social media, we need provably reliable data collection methods, unbiased samples, and valid measures of theoretical constructs of interest. The questions we ask here are:

- Do common data collection methods produce consistent results in terms of the number of user posts recorded and the content of such posts?
- Can we improve the reliability of a data collection effort by combining the results of multiple measurements?
- How should we assess and report the reliability of our data collection efforts?

Methods

Correlation measures the strength of association between two variables and – in ordinal or interval variables – the direction of that association. A reliable data collection method applied twice in order to assess test-retest reliability should produce strongly positively correlated data. We would expect the same of two concurrent measurements with the same objective. The most commonly used correlation metrics are only well defined for numerical, not textual data. Pearson's *rho* product moment correlation gives a correlation value between -1 and 1 for normally distributed ratio or interval variables, while Spearman's *rho* rank correlation coefficient is commonly used for assessing the strength of associations between ordinal variables. Additional metrics have been proposed to assess the reliability of survey instruments, such as Cronbach's *alpha*.

A question remains with respect to calculating correlation values for text, such as when we want to test whether a data collection method has correctly recorded user posts on a social network site. In this section we discuss two methods for producing a correlation metric for text strings. We also propose the use of an information entropy metric to diagnose issues in collected datasets (which can also be used to quantify the improvement in informational content that can be achieved by combining data from multiple measurements).

Normalized edit distance (NED)

A common approach in string matching is to calculate a distance metric between two strings based on the minimum number of operations required to transform one string into the other (Navarro 2011). We used Levenshtein distance, as it is readily available in open source implementations, allows for the three basic operations of addition, deletion and substitution of characters, and can accommodate strings of different lengths. But the fact that edit distance values depend on the length of the strings under comparison makes this unsuitable as a metric for comparison across data where strings may have greatly variable lengths, in which case it would also be difficult to define what values constitute strong or weak correlation. A simple workaround is to normalize edit distance values in the range [0,1]. If δ is the Levenshtein distance between two strings S_1 and S_2, with lengths l_1 and l_2 respectively, taking $l = \max\{l_1, l_2\}$ gives the fowling simple formula for normalized edit distance:

$$NED = (l - \delta)/l$$

The units of the resulting metric are somewhat ill-defined, given that l is measured in characters and δ in operations. Moreover, even if we may be able to normalize values thus, there is no theoretical or analytical connection between the definition and calculation of commonly employed correlation metrics for numerical data (such as Pear-

son's *rho*) and normalized edit distance (NED). This could pose problems when deciding what range of NED values can be interpreted as strong or weak correlation, as is common practice with other correlation metrics. We therefore also consider another method.

Longest common subsequence (LCS)

We first recognize that the problem of string matching is reducible to a sequence of simple dichotomous (binary) decisions: either two characters in a string match, or they do not. A simple approach to string matching, for two strings S_1 and S_2 of the same length l, would be to compare each character in string S_1 to the respective character in the same position in S_2, noting the number of matched characters m. Dividing this number by l would give us an indication of the degree of similarity between the strings.

Since our interest is in matching natural language texts, and the smallest unit of meaning making is the word, we can do the same at the word level and note the number of matching words. Beyond words, meaning is also dependent on the order of words in a sentence. Also, a sequence (of letters or words) can be present in both strings, but phase-shifted. The algorithm we use is a variant of the standard algorithm for the *longest common subsequence* (LCS) problem (Bergroth et al. 2000). We then take the simple ratio

$$LCS = m/l$$

as our string correlation estimate, where m is the length of the longest common subsequence and l again equals $\max\{l_1, l_2\}$. We were able to produce analytical proof that this correlation metric is directly related to Pearson's product moment correlation for numerical data, which has several theoretical and practical benefits. The proof is ommitted here due to space constaints.

Information entropy

A measure of information entropy can be used to assess the amount of useful information in a dataset. For the purposes of comparing entropy across different datasets with imperfect information, we take entropy H for a dataset D of observations taken from a population Π to be:

$$H(D) = -\sum_{i \in \Pi} p(i) \ln\left(p(i)\right)$$

where i is a member of Π, $p(i)$ is the probability that i will be observed in D and moreover, that it is observed correctly (i.e. that the observation i is recorded exactly the same as in the population the observation is taken from, or $i_D = i_\Pi$). This yields:

$$p(i) = p[i \in D \text{ and } i_D = i_\Pi]$$
$$= p[i \in D | i_D = i_\Pi] * p[i_D = i_\Pi]$$

where $p[i \in D | i_D = i_\Pi]$ can be estimated by the relative frequency $f(i) = a(i)/n$ of i in D – assuming D is a sufficiently large random sample of size n or a complete, unbiased crawl of Π, with $a(i)$ being the number of occurences of observation i in D. In cases where i is recorded only once, we take $p[i_D = i_\Pi]$=1, whereas in cases where we may have recorded more than one version of i_D (which is important here as we are concerned with potential errors, duplication, and inconsistencies in collected data), we can take $p[i_D = i_\Pi] = 1/k(i)$, where $k(i)$ is the number of alternative versions of i recorded in the dataset.

Example application of methods

We will now illustrate with an example how the measures we propose can be employed in a data collection task, to assess and report the reliability of data collection methods in a standardized manner, diagnose issues and correct them. In this example taken from past research on social media use, we queried Twitter's Search and Streaming APIs continuously on separate computers during the Greek Indignados, or 'Aganaktismenoi' movement of 2011, for keywords strongly associated with the movement. Moreover, we queried the Search API from two additional locations, on different machines, each connected to a different network. Specifically for reliability testing we selected three keywords and a two-month period, during which we made sure that the data collection tools were fully operational at all times and at all locations, so as to minimize the possibility of errors on our side.

First we provide a summary of the number of entries recorded by each method and at each location. We use the abbreviation "Se" for querying the Search API and "St" for the Streming API. We refer to the different measurement locations as L1, L2 and L3 respectively and denote the use of a method at a specific location with the "@" sign. Table 1 lists the number of entries (tweets) recorded by each measurement. Both methods record the user's Twitter handle (name), the time of the posting, the tweet text, tweet id, and a number of other fields that we will not discuss.

Measurement	Se@L1	Se@L2	Se@L3	St@L3
Entries	66,566	66,850	66,815	124,790

Table 1: Number of entries recorded (raw data)

We observe that querying the Search API yielded similar numbers of entries at each location, while querying the Streaming API seemed to yield about twice as many. This already casts doubt as to the reliability of the data collection. How many tweets were posted during the two-month period under study after all?

We proceed to apply the NED and LCS methods to compute the correlation between tweets in the respective datasets. To do a comparison between two datasets, we must first identify pairs of matching tweets. This can be done by using the tweet id as a unique key, but erring on the side of caution, we also used user name and unix time information

in order to identify unique tweets in each dataset with greater certainty, and report the average of all correlations in Table 2. We provide only one table of correlation estimates, as the NED and LCS methods produced nearly identical estimates. If for an entry present in one dataset we cannot find a matching tweet in the other, we take correlation to be zero.

Average correlation	Se@L1	Se@L2	Se@L3	St@L3
Se@L1	-	0.89	0.89	0.42
Se@L2		-	0.92	0.43
Se@L3			-	0.43
St@L3				-

Table 2: Correlation matrix for raw data (all tweets)

We notice immediately that querying the Search API across different locations yielded highly correlated entries, as expected, but nevertheless, each run yielded somewhat different results. This demonstrates that even when data collection is performed with automated software tools whose instructions are well defined and immutable, the output of the process is *not* deterministic; Results will vary every time the tool is used. We are not always in a position to know this of course for a given data collection tool, unless we run it multiple times in sequence (to assess test-retest reliability), or, in the case of ephemeral streaming data, by running multiple instances of the tool simultaneously, as we did. We also notice in Table 2 that the correlation between the results of querying the Streaming and Search APIs respectively is much lower. This is partly attributable to the observation we made earlier, that St@L3 recorded about twice as many tweets as Se@L3. It would appear that St@L3 was a superior data collection method. To test this we calculate entropy for each dataset and list the values in Table 3.

Measurement	Se@L1	Se@L2	Se@L3	St@L3
Entropy	11.0	11.0	11.0	10.4

Table 3: Entropy of raw data

We notice that entropy values are similar when querying the Search API at all locations, while for St@L3, entropy is slightly lower, in spite of the fact that sample size for St@L3 is almost twice as large. This suggests that many of the additional data points recorded with the help of the Streaming API provide little to no additional information and are in fact, wasteful. Further scrutiny revealed that all datasets (and especially St@L3) contained duplicate entries, with only some of those being exact duplicates. We found edit distance among duplicate entries to be Pareto distributed. Moreover, each dataset contained tweets not recorded in the other datasets. About 20% of all unique recorded tweets were present in either the Search or Streaming API results, but not in both. By combining entries from all datasets we are able to improve the total informational content of our data collection.

Conclusion

We showed how the approach we proposed, in combination with the replication of the data collection procedure and the use of multiple data collection methods, can be used effectively to assess and communicate the reliability of a data collection, as well as to diagnose specific issues before proceeding with data analysis. Neither method produced a complete record of all tweets posted by users, which is consistent with observations in (González-Bailón et al. 2012; Morstatter et al. 2013). We advise other researchers to build such replication into their research designs and report reliability coefficients accordingly.

References

Bergroth, L., Hakonen, H. and Raita, T. A Survey of Longest Common Subsequence Algorithms. In *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE)* (2000), 39-48.

Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G. Network Analysis in the Social Sciences. *Science*, 323, 892-895.

Carmines, E. G. and Zeller, R. A. *Reliability and validity assessment*. Sage, Thousand Oaks, CA, 1979.

Curtis, R. F. and Elton, F. J. Multiple Indicators in Survey Research. *American Journal of Sociology*, 68, 2, 195-204.

Fricker, R. D. and Schonlau, M. Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods*, 14, 4, 347-367.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. Assessing the bias in communication networks sampled from twitter. arXiv (2012).

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of ICWSM* (2013).

Navarro, G. A guided tour to approximate string matching. *ACM Computing Surveys* 33 (1), 31–88.

Rea, L.M. and Parker, R.A. *Designing and Conducting Survey Research*. Wiley, CA, 2005.

Tufekci, Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. arXiv (2014).

Wright, K. Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, 10, 3, article 11.