Online Social Networks and Media 1 (2017) 1–13

Contents lists available at ScienceDirect

Online Social Networks and Media

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

Discovering, assessing, and mitigating data bias in social media

Fred Morstatter*, Huan Liu

Computer Science and Engineering, Arizona State University, Tempe, AZ, USA

ARTICLE INFO

Article history:
Received 1 December 2016
Accepted 17 January 2017
Available online 30 March 2017

Keywords:
Social data bias
Data collection
Data collection bias
Social media mining
Twitter
Machine learning
Data mining

ABSTRACT

Social media has generated a wealth of data. Billions of people tweet, sharing, post, and discuss everyday. Due to this increased activity, social media platforms provide new opportunities for research about human behavior, information diffusion, and influence propagation at a scale that is otherwise impossible. Social media data is a new treasure trove for data mining and predictive analytics. Since social media data differs from conventional data, it is imperative to study its unique characteristics. This work investigates data collection bias associated with social media. In particular, we propose computational methods to assess if there is bias due to the way a social media site makes its data available, to detect bias from data samples without access to the full data, and to mitigate bias by designing data collection strategies that maximize coverage to minimize bias. We also present a new kind of data bias stemming from API attacks with both algorithms, data, and validation results. This work demonstrates how some characteristics of social media data can be extensively studied and verified and how corresponding intervention mechanisms can be designed to overcome negative effects. The methods and findings of this work could be helpful in studying different characteristics of social media data.

1. Introduction

Social media is an important outlet to understanding human activity. Over the last few years many social media sites have given users a way to express their interests, friendships, and behavior in an online setting. Because of their ubiquity, these platforms have been critical in many global events. During the Arab Spring protests, these platforms helped protesters to organize. Across several natural disasters such as Hurricane Sandy, earthquakes, typhoons, and floods, social media has been used both by the affected to request assistance as well as by humanitarian aid agencies to spread information about critical aid resources. More generally, social media is used by everyday people to discuss current events, and their day-to-day activities. There are several sites with hundreds of millions of users, and a few sites with billions of users, all sharing, posting, and discussing what they see around them.

Noticing the richness, extent, scale, and dynamic nature of social media data, researchers welcome the new opportunities to use social data to answer questions regarding human behavior. Though not all social media sites provide data for research purposes, some sites do provide mechanisms through which researchers can obtain a sample of data to conduct their research. One example is Twitter, a microblogging site where users exchange short, 140-character messages called "tweets." Ranking as the 8th most popular site in the world by the Alexa rank in August of 2016,[1] the site boasts 313 million monthly users publishing 500 million tweets per day. Twitter's platform for rapid communication is a vital communication platform in recent events including Hurricane Sandy,[2] the Arab Spring [1], and several political campaigns [2,3]. As a result, Twitter's data has been coveted by both computer and social scientists to better understand human behavior and dynamics. Because of its open nature with sharing data as well as the richness and size of the data generated on Twitter, many research projects use Twitter data to understand human behavior online. This has led to Twitter being called the "fruit fly", or model organism, for computational social sciences research [4].

While Twitter is an amazing resource for research on social media, there may exist bias during data collection of which researchers should be aware in their study, i.e., whether a representative dataset is obtained for planned computational social science research. If the goal of computational social science is to study society at scale, then the data we study must provide an accurate reflection of society. More specifically, we study whether or not the sampled data that researchers often use for their research is rep-

* Corresponding author.
E-mail addresses: fred.morstatter@asu.edu (F. Morstatter), huan.liu@asu.edu (H. Liu).

[1] http://www.alexa.com/topsites .

[2] http://www.nytimes.com/interactive/2012/10/28/nyregion/hurricane-sandy.html .

http://dx.doi.org/10.1016/j.osnem.2017.01.001
2468-6964/

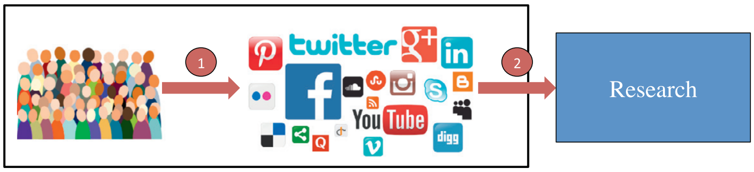

Fig. 1. Overview of the process: human behavior becomes data upon which research is conducted. Humans generate data on social platforms ("1") which is then collected by researchers in order to answer questions about their behavior ("2"). At both steps, there is a potential for bias. In this work, we focus on API bias, denoted by "2" in the figure.

resentative of the full, unsampled data on Twitter, Firehose data. If there is bias, we ask if we could detect bias under normal circumstances without the aid of Firehose. In this work, we focus on bias that arises from sampling strategies on social media and potential sources of data bias, and present ways of detecting and mitigating bias in order to draw credible conclusions from sampled and limited data.

2. Related work

This section consists of two parts that are relevant to this study. The first part is related to existing work on data collection bias and the second part is about the technical details of data gathering mechanisms available at Twitter.

2.1. Data collection bias

First, users on a site can introduce bias into a dataset. This corresponds with arrow "1" of Fig. 1. This is often done unintentionally by the user base of the site. For example, Twitter's user base consists mostly of young users [5]. Thus, any body of tweets is likely to have a very different age distribution than the age distribution within a country. Partly due to this, we cannot generalize conclusions made from Twitter data without taking these differences into consideration. These demographic biases have been studied previously. For example, in [6] the authors discover key demographic dimensions in which Twitter demographics differ from the demographics of the real world. Addressing the concern about generalizability explicitly, [3] found that using Twitter alone to predict elections did no better than random chance. The authors attribute this poor performance to the demographic makeup of the site.

While the body of genuine users on the site can present bias to those studying the aggregate of their posts, malicious users can also introduce bias into the site. Bots, software-controlled accounts, can work in tandem to change the statistics of the site. They can cause a topic to trend [7] or they can misrepresent already trending topics.[3] Bots can also be used to follow specific user accounts, making those accounts appear more prominent than they actually are [8].

Malicious users are not restricted to bots. In fact, there are many humans who coordinate to damage the reputability of social media sites. One way that this is done is through "crowdturfing," [9] where humans are hired to perform specific tasks. These tasks often take the form of fake reviews [10], where coordinators organize negative reviews of competing products or positive reviews of their own. Additionally, a new phenomenon of "shills" has appeared on social media. This is where well-trained people disguise themselves on social media, usually to address negative information of a campaign.[4]

In this work, we focus on the general area of bias in social media data collection. For this reason we largely focus on arrow "2" of Fig. 1 in the rest of the paper. However, we would be remiss to omit some important extensions to this area. For example, some work discusses how one can use expert sampling to surpass the randomness of data collection through APIs [11]. In another comparison, research has been carried out to compare Twitter's Search and Streaming APIs [12]. This allows a deeper understanding of which dataset to use when the Firehose is not available.

2.2. Overview of Twitter's API mechanisms

In this work, we use Twitter's APIs as an example of biased data. We provide a brief introduction to the two APIs studied in this work. Twitter provides several APIs which allow researchers and practitioners to collect data to answer their particular research question. The "Twitter Streaming API"[5] is a capability provided by Twitter that allows anyone to retrieve at most a 1% sample of all the data by providing parameters. The sample will return at most 1% of all the tweets produced on Twitter at a given time. Once the number of tweets matching the given parameters eclipses 1% of all the tweets on Twitter, Twitter will begin to sample the data returned to the user. The methods that Twitter employs to sample this data is currently unknown.

2.2.1. The Twitter Firehose

One way to overcome the 1% limitation is to use the Twitter Firehose – a feed provided by Twitter that allows access to 100% of all public tweets. However, the Firehose data is very costly. Another drawback is the sheer amount of resources required to retain the Firehose data (servers, network availability, and disk space). Consequently, researchers as well as decision makers in companies and government institutions are forced to decide between two versions of the API: the freely-available but limited Streaming, and the very expensive but comprehensive Firehose version. To the best of our knowledge, no research has been done to assist those researchers and decision makers by answering the following: how does the use of the Streaming API affect common measures and metrics performed on the data? In this article, we answer this question from different perspectives.

2.2.2. Streaming API

Using the Streaming API, we can search for keywords, hashtags, user IDs, and geographic bounding boxes simultaneously. The *filter* API endpoint[6] facilitates this search and provides a continuous stream of Tweets matching the search criteria. The limitation of the Streaming API is that it will return, at most, 1% of all of the tweets on Twitter. When a query stays below the 1%, then the Streaming API can return all of the tweets pertaining to that query. Once the volume of tweets surpasses 1% of all of the tweets on Twitter then the results will be sampled. How this sampling process is carried out is not published by Twitter.

[3] http://krebsonsecurity.com/2011/12/twitter-bots-drown-out-anti-kremlin-tweets/.

[4] http://www.thedailybeast.com/articles/2016/04/21/hillary-pac-spends-1-million-to-correct-commenters-on-reddit-and-facebook.html .

[5] https://dev.twitter.com/docs/streaming-apis .

[6] https://dev.twitter.com/streaming/reference/post/statuses/filter .