Social Media Data: Quantitative Analysis

Foundation Entries

SAGE Research Methods Foundations

By: Subhayan Mukerjee & Sandra González-Bailón
Published:2019
Length: 10,000 Words
DOI: http://dx.doi.org/10.4135/9781526421036
Methods: Social Media Data: Quantitative Analysis
Online ISBN: 9781526421036
Disciplines: Communication and Media Studies, Political Science and International Relations, Psychology, Sociology
Access Date: February 19, 2020
Publishing Company: SAGE Publications Ltd
City: London

Abstract

This entry describes research analyzing social media with quantitative techniques. Recent studies using social media data are first situated in one of four categories as defined by the research design (observational or experimental) and the object of analysis (network structure or content). Several examples from the field of political communication are used to illustrate quantitative research designs based on social media data. These examples review research in areas from political participation and selective exposure to news consumption and the diffusion of misinformation. The entry concludes with a section on the limitations of using social media data by identifying some of the ethical concerns that emerge in this research domain.

Introduction

Quantitative social science relied, for much of its history, on two methodological tools: experiments and surveys. These methods involve asking people questions about their relationships, preferences, and choices or observing their actions in highly controlled environments. The advent of the Internet and the subsequent rise of social media platforms opened a new methodological era in social science research. These technologies made it possible to observe people's behavior and actions in nonexperimental settings that mimicked more closely (although not perfectly) the real world. Of course, the use of observational data for social research predates the Internet. Research on audience behavior, for instance, has long analyzed behavioral trails such as viewership patterns as measured by television meters. But never before have observational data been available at the scale and resolution that it is available now (Gonzaález-Bailoón, 2017). As a result today, people create new data daily as a by-product of their interactions with technology.

Social media offer a prominent example of this data revolution. Since the turn of the 21st century, platforms such as Facebook, Twitter, YouTube, and Instagram have seen unprecedented growth in their user base. According to a 2018 report from the Pew Research Center, more than two thirds of Americans use Facebook and YouTube, more than one third use Instagram, and close to one quarter use Twitter (Smith & Anderson, 2018). These platforms are also among the most widely visited websites globally, with YouTube and Facebook consistently ranking among the top three domains. Moreover, the rise of social media has also engendered a new culture of information diffusion—one that is driven by attempts to make content go "viral" (Goel, Anderson, Hofman, & Watts, 2015).

The overwhelming popularity of social media platforms and the rise of the digital advertising industry have made the epithet "data is the new oil" an often-invoked cliché. There are many downsides to such a media environment (and some of them will be dealt in greater detail later in the entry) but the possibility to extract user data at scale has, undoubtedly, prompted a paradigm shift in the kind of social science research that can be conducted today. Social media have become one of the most popular data sources for quantitative researchers, offering both a tool and a lens to look into social behavior to better understand processes at the individual, collective, and social levels (Gonzaález-Bailoón, 2017).

This entry first describes the two types of research design that social media data help implement (i.e., observational and experimental) and then talks about the main areas of inquiry developed using these data, giving specific examples of research that focus on the structure of interactions or on the content of the messages exchanged. The entry concludes with an assessment of the limitations of social media data and a discussion of the ethical issues that arise in this line of research.

Types of Research: Observational and Experimental

Quantitative research using social media data can take one of two forms, each of which predates most digital technologies. The first is observational research, which entails the systematic reconstruction of actions and behavior as observed and measured in natural settings. This approach to understanding the social world aims to supplement, or at times even supplant, the information that surveys provide. Unlike surveys, observational data do not rely on subjective responses to specific questions but on what people do when engaging in interactions with other people or with an environment. Observational research thus engages in passive data collection and is consequently less intrusive. It is also often superior at reducing measurement error than surveys, which can generate recall problems and variation in respondents' interpretation of the same questions. One example of observational research is a study of diffusion in Twitter in which researchers find a correlation between positive content and retweets received (Bakshy, Hofman, Watts, & Mason, 2011). In another example, researchers analyze a billion diffusion events on Twitter and measure how many of those events count as "viral" (Goel et al., 2015). A third example looks at protest activity on Twitter to reverse engineer the growth of a political movement as it evolved online (González-Bailón, Borge-Holthoefer, Rivero, & Moreno, 2011). Another example analyzes how millions of Facebook users interact with news articles to conclude that personal networks limit exposure to cross-partisan political content more than algorithmic curation does (Bakshy, Messing, & Adamic, 2015). These studies are discussed in greater length in later sections.

The second type of research that social media enables makes use of field experiments. In these experiments, researchers carefully administer interventions to different groups of users on a social media platform to better understand the causal mechanisms driving their behavior. Examples of this type of research include a field experiment on Twitter in which researchers paid participants to follow a bot which regularly tweeted political (ideologically opposing) messages (Bail et al., 2018). At the end of the experiment, the researchers surveyed participants about their political beliefs, and they found that, contrary to conventional wisdom, being exposed to ideologically opposing messages made participants more polarized. In another study, researchers conducted a randomized experiment on Facebook to measure social influence (Aral & Walker, 2012). After a set of users adopted a new product, notifications of their adoption were sent to a random set of their friends in order to determine whether those exposed to this information were more likely to also adopt the product than those who were not notified. Another group of researchers used a field experiment on Facebook to study social influence and political mobilization (Bond et al., 2012). This study found that exposure to information on the voting behavior of friends had a direct positive influence on political self-expression, information seeking, and voting. Elsewhere, researchers conducted a massive field experiment on multiple social media sites to unpack the censorship dynamics at play in China (King, Pan, & Roberts, 2014). Yet another example of an experimental research design involved a bot intervention in Twitter to reduce racist slurs (Munger, 2017). The following sections discuss each of these studies in more detail.

Accessing Social Media Data

In the examples that follow, and in studies that analyze social media in general, the first step is always the collection of data. The use of application programming interfaces (APIs) is the most popular and effective way to collect data sets tracking social media activity. APIs are sets of programming functions that afford people the ability to gain access to data stored in the servers maintained by a platform or a service, in a manner the platform or the service deems appropriate and in line with their terms of service (ToS). Most social media platforms like Twitter and Facebook maintain formal APIs for various different programming languages; however, because the protocols of access are determined by their own business priorities, they can change from time to time.

Outside of the use of formal APIs, another means of obtaining social media data is web scraping, a data collection technique that automatically extracts information from web pages including those of social media platforms. This approach to data collection is often explicitly prohibited in the ToS, which means that researchers can be subject to litigation if they violate those terms.

Finally, other means of gaining access to social media data include the use of third-party services like CrowdTangle (for Facebook) or Twitonomy (for Twitter). These commercial services allow the use a graphical interface to enter the specifications of the data needed. This interface connects to the formal APIs in the back end that fetches the required results.

The data extracted through the use of APIs or scraping techniques are usually obtained in JavaScript Object Notation (JSON) format. JSON is a structured text file format that allows users to parse and transmit human readable data. Once the data are obtained in JSON, it can be converted into more easily programmable formats like tables and data sets or stored in a database for future analysis. Sometimes, the data are stored in a comma separated values (CSV) format, which is also a simple text file where data are recorded in a tabular format, with entries separated by commas.

Research on Structure: Networks, Interaction, and Diffusion

The data generated by social media (and collected through the means discussed in the previous section) is a by-product of user behavior on the platforms. Users generate digital trails as they navigate online networks and as they interact with content or other users. Since social media platforms are “always on” (Salganik, 2017, p. 21), they serve as a repository of social activity that is continuously updated in real time. This activity includes changes in networks of interaction or in the frequency with which the ties in those networks are activated—for instance, to spread information or engage in conversation. This “behavioral residue” offers interesting new insights into social dynamics and the underlying structure of interactions. The analysis of digital data can, in fact, shed light on different aspects of social life, from the way in which individuals act to how organizations, collectives, or even societies behave. As with any type of empirical research, however, trace data alone are not enough to advance our understanding of social life. The choices made when transforming data into meaningful patterns are also a crucial element of research design. The following two sections focus attention on those choices and, in particular, on how structures of interaction and dynamics of social influence and diffusion can be analyzed.

Social Media Data as a Proxy to Social Networks

Social media data allows for the reconstruction of different types of networks: Some track communication patterns, while others track interpersonal relationships or friendship ties. Social media platforms offer data that can be used to build these different types of structures. These data serve as the foundation for many studies looking at online interactions or at how ideas diffuse in a population. At the heart of this approach is the concept of the network, and formal mathematical tools to analyze those networks and what they reveal about social structure. For instance, Facebook interactions rely on reciprocal connections: Two users need to agree that a connection exists for them to form a tie. Researchers have found that the structure of those online relationships reproduces some of the properties long theorized to characterize offline social networks—in particular, small world features (Ugander, Karrer, Backstrom, & Marlow, 2013). These features suggest that any two random nodes in the network are at a short distance from each other, usually four to six steps away or, to put it differently, separated by a few intermediaries. This feature is one of the reasons why social networks are often so efficient at diffusing information: their structure has many shortcuts that allow information to flow fast (Watts, 2004). But data collected from Facebook can also be used to infer other types of connections: For example, researchers can look at who sends messages to whom, and whether those communication ties are reciprocated or asymmetrical. As the definition of what counts as a tie gets stricter, the network becomes smaller and sparser. Likewise, a platform like Twitter can be used to build different types of networks: There is the basic structure of who follows whom, but there are also the additional network layers of who sends messages to whom or who retweets content produced by others. Each of these networks has different structural properties, and they can be compared to assess how political campaigns unfold in the preexisting ties of social media (González-Bailón & Wang, 2016) or whether polarization in political discussions is more or less pronounced across modes of interaction (Conover, Ratkiewicz, & Francisco, 2011).

The nodes in these networks are user accounts, most frequently managed by individuals but also by organizations and even nonhumans (as when accounts are automated with code in the form of bots). One question researchers often ask of these data is how the structure of the networks shapes exposure to content or who is in a better position to disseminate information. For example, if the retweet network shows evidence of ideological polarization, it means that diffusion chains are trapped within groups of like-minded people; but if the network formed by mentions does not exhibit polarization, it means that those same users might actually be engaging in cross-cutting discussions, or at least, exposure. Likewise, if some users are bridging two clusters in the network (clusters that would otherwise be separated), these users can gain influence by controlling information flows. Research that pays attention to bots and their role in the coordination of propaganda campaigns aims to determine precisely, if these bots hold strategically important positions to control the flow of information.

Information Diffusion and Social Influence

Social networks act as conduits of interpersonal influence and information diffusion. By tracking how content disseminates through those networks, it is possible to unpack how a critical mass is formed or how transmission chains involving a specific type of information grow (Gonzaález-Bailoón, 2017). The metaphor that information can spread like a virus has become a central tenet of the digital age. Social media platforms have enabled faster avenues for diffusion compared to those that existed during the broadcast era, which was characterized by much more centralized communication structures (Welles & González-Bailón, 2018). The properties of online networks, however, do not guarantee that every diffusion event will trigger cascading behavior or large diffusion chains; most instances of diffusion, in fact, barely go beyond the initiating person (Goel et al., 2015). This begs the question of how much the content of a message shapes the probability that the message will diffuse widely by making people more likely to react to it and decide to spread it. Researchers studying how emotion spreads on Twitter show that emotionally charged tweets tend to be retweeted more often (and therefore spread more widely) than emotionally neutral tweets (Stieglitz & Dang-Xuan, 2013). Likewise, research with Facebook data shows that diffusion is more likely to unfold when people are influenced by friends from different social circles, a process that reinforces the impact of exposure to information (Ugander, Backstrom, Marlow, & Kleinberg, 2012). Factors like the emotional tone of messages or whether the messages received from different people contain the same information require analyzing the content of social media communication—the topic of the following section. Other examples of research on diffusion are described in the “Propaganda and Misinformation” section later in this entry.

Research on Content: Sentiment, Topics,

and Images

In addition to providing insights on the structure of interactions and diffusion trails, social media platforms are rich sources of data on textual communication and visual content. These data can offer valuable information on public opinion trends, framing dynamics, or shifts in collective attention. It can also shed light on discursive dynamics, the quality of deliberation, or civility in political talk. Depending on the substantive object of interest, research on the content generated by users on social media can focus on text or images, as the following two sections explain.

Sentiment and Civility

Social media is a repository of public expression. Owing to their popularity, platforms like Twitter and Facebook have become perhaps the largest known archives of public opinion in the world. Consequently, they have proven to be extremely useful platforms for conducting research on opinion trends. One prominent aspect of those trends is the nature of the opinions expressed and, in particular, the valence or sentiment expressed in those opinion (i.e., favorable or unfavorable, positive or negative). Tools originally developed in a subfield of computer science known as natural language processing—which builds algorithms to help computers understand the nuances of human linguistic expression—have seen wide adoption in the social sciences in recent years, with many applications in fields like political communication.

Natural language processing (NLP) can be used in a variety of contexts to understand the emotional tone of messages, ideological leaning, or the level of civility in what people say on social media. Various implementations of standard NLP functions can be found as packages in popular programming languages such as R or Python, and they vary greatly in the degree of complexity of what they can achieve.

In one study, for instance, researchers connect expressed sentiment measured from tweets with public opinion measures gathered through surveys polling a representative sample of the U.S. population (O'Connor, Balasubramanyan, Routledge, & Smith, 2010). To do this, they analyzed over 1 billion tweets spanning a 2-year period (2008–2009). They first classified the tweets in three categories: messages expressing consumer confidence (e.g., containing keywords like *economy, jobs, job*), messages expressing presidential approval (e.g., with keywords associated to the then U.S. president Barack Obama), and messages related to the 2008 elections (e.g., with keywords related to the two U.S. party candidates). They then defined the daily sentiment score of the topic as the ratio between the number of "positive" tweets in that topic to the number of "negative" tweets in that day. A tweet was deemed "positive" or "negative" depending on whether it contained a predetermined set of "positive" or "negative" words as defined by a dictionary of words (e.g., a subjectivity lexicon known as *OpinionFinder*). Analyzing these data, they found that their basic sentiment analysis was able to achieve a high correlation with survey measures of public opinion.

Similar examples include the analysis of tweets to extrapolate measures of happiness at the population level (Dodds & Danforth, 2010) and to characterize the civility of political talk (Jaidka, Zhou, & Lelkes, 2018). In another study (discussed in a later section), researchers at Facebook ran a field experiment to see how sentiment spread through the users of the platform by curating their news feeds and increasing the number of positive or negative posts they were exposed to (Kramer, Guillory, & Hancock, 2014).

Topics and Visual Content

Another text mining technique that has gained significant traction in the social sciences is that of *topic modeling*. Topic modeling refers to a class of statistical techniques that are used for discovering conceptual topics in a set of documents or texts. One of the most widely used implementations is known as Latent Dirichlet Association, or LDA (Blei, Ng, & Jordan, 2003). Researchers have used this technique to identify and compare the topics discovered in a representative sample of tweets with the topics found in news articles obtained from *The New York Times* during the same time period (Zhao et al., 2011). They found that while the topics largely correlated, Twitter and *The New York Times* disagreed when it came to the order of importance or popularity of certain topics.

Other approaches to identifying topical clusters include the use of networks. Studies have shown that the visualization of retweet and mention networks on Twitter can be used to identify thematically homogenous publics and counterpublics (Jackson & Welles, 2015). This analysis can then be used to map the ebb and flow of narratives and counternarratives on online platforms.

The analysis of visual content shifts attention from text to images. One study, for example, looks at images exchanged in social media during political mobilizations to determine whether affectively loaded images elicit greater participation and engagement (Casas & Williams, 2018). The study finds that tweets with images that invoke enthusiasm and fear are more likely to be retweeted. Another example looks at food images posted in Instagram to determine what affects their liking and sharing and explore implications for the marketing of healthy eating (Peng & Jemmott, 2018). This line of research, however, is still incipient and likely to grow substantially in the next few years.

Applications: Social Media and the Political Process

One prominent area in which the various approaches described thus far have been applied relates to politics. This section highlights some of the areas in political communication in which quantitative social media research has made progress in recent years, offering several examples that illustrate the state of the art.

Social and Political Homophily

In sociology, social media data are often used for testing and informing new theoretical frameworks and for understanding structural dynamics of social relationships. Both Twitter and Facebook, for instance, offer valuable data about users' networks that help us understand how individuals form ties and make social connections. Homophily, or the tendency of similar individuals to interact more with each other, is an idea that has long permeated sociological thinking. Social media platforms provide researchers with a testing ground of how such dynamics operate. Follower-following networks on Twitter, for example, have been used to study how people who are similar along certain dimensions (they are from the same region, or have the same gender, or similar ethnicities, or hold similar professions) tend to connect more, show similar interaction patterns, leave similar behavioral traces, and show interest in similar topics (De Choudhury, 2011). A similar study conducted with international Facebook friend networks also found that connections are largely determined by physical proximity, cultural closeness, and actual patterns of communication between countries like migration (Barnett & Benefield, 2017).

Even outside the realm of descriptive studies, the concept of homophily has often been used to build predictive models that can help provide sharper, and more interesting insights into social processes. In a 2011 study, researchers used the features of a user's social network (e.g., network structure and the content of their feeds) to improve inferences on the political ideology of users (Pennacchiotti & Popescu, 2011). Yet another study found evidence that the content generated by a user's friend network (e.g., words tweeted, interactions) are better predictors of the user's age, gender, and political affiliation than the content generated by the user him- or herself (Zamal, Liu, & Ruths, 2012).

Similar intuitions have also motivated political scientists and communication researchers to identify the factors that explain the formation of political echo chambers online. The concept of political homophily has seen widespread use in studies that have sought to infer the political ideologies of social media users from their social networks. This line of research typically builds on the idea that political ideology is a crucial determinant of social behavior and is, in turn, a good determinant of social media activity. This type of data has helped researchers go beyond the use of roll-call votes or bill cosponsorship records to reliably estimate ideology of not just political elites but also ordinary social media users. Researchers have shown for instance, how the structure of Twitter conversations (reply and retweet patterns) can be accurately used to predict a user's ideology (Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer, 2011). The researchers used these predicted ideologies in conjunction with a network science technique known as community detection to find two communities of like-minded users based in the United States, offering evidence of polarization in the dissemination of political information. Another study used the known ideologies of politicians in different countries to infer the ideology of common users following these elites in Twitter (Barberá, 2015). The results obtained by this method not only correlated well with existing methods to infer ideology but has also the added benefit of being able to infer ideology for a much larger sample of users than traditional surveys allow (although issues of representativeness remain).

Polarization and Selective Exposure

Political communication researchers build on methods to infer political ideologies to apply to the analysis of social media theories of partisan selective exposure that were devised to study older forms of media. These studies are motivated by the prominence of social media platforms in today's political process because they continue to remain one of the most important channels to share and consume political information.

The Barberá study already mentioned (2015) analyzed partisan divides in a retweet network involving 150 million messages. The author used retweets as the unit of analysis because retweets are a useful proxy for endorsement: It is expected that people retweet ideologically congruent tweets more than ideologically conflicting tweets. The results of this study show that this is indeed the case. While retweets about nonpolitical issues (e.g., sports) do not show any significant partisan selectivity, retweets about political issues do. Moreover, the researcher also found an asymmetry in the nature of the partisan divide: Consistent with psychological theory, liberals were more likely to retweet conservative tweets than conservatives were likely to retweet liberal tweets.

An earlier study that tried to understand the extent of ideological fragmentation on Twitter first categorizes users based on the linguistic content of their tweets; it then looks at the political homophily of the networks these users form (Colleoni, Rozza, & Arvidsson, 2014). These researchers found (counterintuitively) that while the Democrat network was more homophilic than the Republican network, the subset of the Republican network that follows official Republican accounts is significantly more homophilic than the network of Democrats who follow official Democratic accounts.

Facebook has perhaps provided even greater scope to researchers studying partisan selective exposure owing to its massive user base and general popularity. One study on Facebook tries to determine the extent to which users share news articles (An, Quercia, & Crowcroft, 2013). The researchers found that partisans do predominantly share like-minded articles and avoid conflicting ones. The mechanism underpinning this pattern is similar to political homophily: Political motivations determine what information people seek or are willing to share more. Another determinant, however, is the nature of the platform's algorithmic curation of users' feed—in other words, what does the social media platform want its users to see? It is the simultaneous operation of social and algorithmic dynamics that dictates who gets exposed to what information on social media. A study conducted in 2015 addressed this issue (Bakshy et al., 2015). Researchers used a large data set of Facebook users with self-reported ideologies, and thousands of hard news URLs shared over a 6-month period. They first assessed the political slant of the news stories based on the sharing patterns of the users. They then came up with estimates of the extent to which the political homogeneity of users' friend networks limited exposure to cross-cutting ideological news. Finally, they controlled for the ranking of news stories that users see in their Facebook feeds to conclude that while the news feed does allow users to choose to select ideologically diverse news stories, users tend to click more on those URLs that are more aligned with their ideology.

Another study looks at news consumption patterns on Facebook during the 2016 UK Brexit referendum to assess the extent of selective exposure during a period of particularly heightened political division (Del Vicario, Zollo, Caldarelli, Scala, & Quattrociocchi, 2017). This study makes use of network science techniques to analyze exposure to news sources. The researchers downloaded all the posts made on Facebook by British news outlets in the months leading up to the referendum, including user likes and comments. Then, they built a network of user overlap—where the nodes are the media sources and the edges represent the number of users who liked posts from any pair of sources. The researchers found that network of news sources could be clearly divided into two communities of users. In addition, they used natural language processing techniques to analyze the topics discussed in the news posts, finding that the two communities largely discussed the same topics. In other words, neither community preferred any topic to any other—thus revealing that selective exposure to Brexit information was not driven by interest in different topics. At the same time, however, they found differences in the sentiment with which those topics were covered: One community of news sources projected some of the topics in more favorable light, while the other community projected some of the other topics in more favorable light.

A related line of inquiry relates to the effects that information consumption on social media has on political polarization. Much has been written about how social media exacerbates the existing partisan divides by reinforcing political beliefs and cutting off partisans from opposing perspectives (Sunstein, 2017); other researchers contest these claims (see Guess, Lyons, Nyhan, & Reifler, 2018 for a review). A recent study puts this hypothesis on trial by conducting a field experiment (Bail et al., 2018). In this study, the researchers recruited self-reported Republicans and Democrats who use Twitter frequently and surveyed them about their attitudes on various political issues, before paying them to follow Twitter bots for 1 month. Democrats in the treatment groups were instructed to follow a bot that tweeted pro-Republican messages, and vice versa: Republicans were instructed to follow a pro-Democrat bot; while in the control group, neither Democrats nor Republicans were asked to follow any particular account. At the end of the month, the researchers surveyed the participants again and found that those in the treatment group held more extreme positions on the same issue (and in the same direction). They concluded that social media need not increase political polarization by just reinforcing beliefs but by exposing people to counterattitudinal information as well, triggering a backfiring effect.

Political Participation and Mobilization

There is a rich literature in social science research that tries to understand how individuals use networks of social connections to increase social capital and to mobilize. Social media platforms have given researchers better ways to measure how these dynamics play out in the real world which, in turn, has been used to enrich existing theories of political participation.

Social networking sites offer a “new space for political behavior” (Bode et al., 2014). Research shows consistent evidence that online news consumption increases civic awareness, which in turn affects political engagement (Boulianne, 2016). Looking for information on social media websites is a significant predictor of not just people’s social capital but also their “civic and political participatory behaviors, online and offline” (Gil de Zúñiga, Jung, & Valenzuela, 2012). Social movements in the digital era rely largely on online networks to disseminate information and mobilize. A meta-analysis of studies using self-reported social media use also concludes that “social media use generally has a positive relationship with engagement and its three sub-categories, that is, social capital, civic engagement, and political participation” (Skoric, Zhu, Goh, & Pang, 2016).

The Arab spring in 2011 marked a turning point for research into the potential of social media in facilitating social movements. Other social movements followed: from the “kitchenware revolution” against the “banksters” in Iceland, to the *Indignados* in Spain; from the #Occupy movements across the world, to the #SaveTheInternet movement in India. The use of online platforms like Twitter not only facilitated the mobilization of a critical mass, it also engendered new tactics like hashtag activism (e.g., #Ferguson, see Bonilla & Rosa, 2015). It also gave researchers access to real-time data of how these movements emerge and evolve online, information that can be fed back into the process to inform mobilization strategies.

A large body of literature has emerged in recent years offering many insights into online social movements. One of these studies uses comparative content analysis of manually coded tweets collected during Occupy Wall Street, the *Indignados*, and the anti-austerity movements in Greece (Theocharis, Lowe, van Deth, & García-Albacete, 2015). The study finds that most of the messages were used for political discussion and for the communication of protest information and very few messages called directly for mobilization and protest coordination. Another study scraped Twitter data using the #Ferguson hashtag right after Michael Brown was shot and killed by a police officer in Ferguson, MO, to try and understand how the online discourse about the killing moved to mainstream media (Jackson & Welles, 2015). The researchers analyze how Twitter’s architecture shaped the “discursive strategies” of those actively tweeting in the Ferguson networked public. In order to do this, they used a mixed-methods approach, combining network analysis of the #Ferguson network to first identify the emergent elites in the #Ferguson network, and then qualitatively analyzing the tweets of these elites to understand their framing strategies. The study found that as the online protest unfolded over the week following the killing, opinion leaders (“crowd-sourced elites”) emerged in the immediate aftermath—with the rest of the network heavily retweeting these chosen few in the first few days. Over the week, however, the retweet-mention ratio decreased continuously—although the ideological discourse largely remained consistent and with overwhelming focus on the critique of state violence.

In another study, researchers studied the core–periphery dynamics that characterize protest movements using three different Twitter data sets (Barberá et al., 2015). To formalize the research question, the study applies network science techniques to identify layers of connectivity in retweet networks associated to political protests. The researchers found that the core in these networks represented a dedicated minority of activists, but the visibility of their actions relied on the numerous people in the periphery: These users were not very active on an individual basis, but collectively, they allowed the movements to gain traction and global visibility.

An earlier study looked at the recruitment patterns in Twitter for a political mobilization in Spain and found evidence of social influence and contagion dynamics (González-Bailón et al., 2011). The researchers first collected tweets over a 1-month period using 70 representative hashtags of the protest. Then, they examined the temporal trends in the Twitter networks to identify the “early participants” in the protests and found that they were randomly distributed throughout the network. However, a temporal analysis also revealed bursts of recruitment activity, which resulted from the presence of highly connected, core users who, having followed some of the early participants, were exposed to their protest messages and retweeted them to their broader audiences. These “bursts” of recruitment activity hold the key for the fast growth of online protests.

Social media facilitates not only protests and social mobilization but also political participation. One study that demonstrates how relies on a massive field experiment on Facebook (Bond et al., 2012). In this study, the researchers conducted a randomized controlled trial on all the users of Facebook who were more than 18 years of age, based in the United States, and accessed Facebook on the day of the U.S. Congressional elections (N = 61 million). Each person was randomly assigned to one of three conditions. In one, the “social message” condition, users saw the following: a message urging them to vote; a link to find local polling places; a counter of the number of Facebook users who had reported voting, and some faces of their friends who had reported voting; and a button so that they could themselves report their voting. In another condition, the “informational message” condition, users just saw the counter, the message, the link to find polling places, and the button—but not the faces of their friends. The third condition was the control group, where users saw none of the pieces of information in the other two groups. The researchers then compared user behavior across the three groups—that is, whether they clicked on the button (a measure of political self-expression), whether they clicked on the link to view polling places (a measure of desire to seek political information), and whether they actually voted (a measure of political participation). They found that users in the informational message condition were significantly more likely than users in the control group to self-express, seek political information, and actually go out and vote. Moreover, users in the social information group were even more likely to do all of these. The researchers further looked at the effects of these treatments on the users’ friends and found evidence that the observed per-friend treatment effects increased as the strength of the tie between the friend and the user increased. They concluded that online political mobilization works not just directly but also through social influence and that online messages can in general be used to influence offline behavior.

Another study sought to understand social media effects by looking for causal evidence of the classic agenda setting theory through an experiment that was (partially) fielded on Twitter (King, Schneer, & White, 2017). The researchers recruited a number of small media outlets and instructed them to publish articles on specific issues on randomly chosen dates and then assessed the impact that those articles had on the Twitter discussion networks. The researchers found evidence that their intervention increased the prominence of these issues in social media, thereby lending credence to the agenda-setting function of mass media.

Propaganda and Misinformation

The role of social media platforms in the diffusion of misinformation and as tools for political propaganda has become a growing concern in the second decade of the 21st century. To address this concern, one study analyzed the sharing patterns of hundreds of thousands of news stores over an 11-year period on Twitter (Vosoughi, Roy, & Aral, 2018). The researchers found that, irrespective of the category of the news story, false news diffuse faster, farther, and deeper than true stories. To do this, the researchers first sampled all Twitter rumors that had been investigated by six different fact-checking organizations and then obtained all the English language replies that each of them had received on Twitter. They then built the cascade networks and analyzed the size, the depth, the breadth, and other structural aspects of the diffusion patterns of these cascades.

Another study conducted by Nir Grinberg and colleagues (2019) found that exposure to fake news and misinformation on Twitter in the lead-up to the 2016 U.S. presidential elections was not widespread and concentrated only among a very small set of U.S. voters. The researchers accomplished this by linking a sample of U.S. voters to Twitter accounts and analyzing the extent to which they posted content from news websites that had been verified to be of questionable repute by fact-checkers, journalists, and academics. While overall exposure to fake news was low, people who were more politically engaged, older, and conservative tended to be more exposed to fake news than other users.

A related area of social media research is in the sphere of automated propaganda by the use of bots. An example is a study on bot activity during the 2017 Catalan referendum for independence (Stella, Ferrara, & De Domenico, 2018). By studying the Twitter interaction network between humans and bots, the researchers found that while bots did exist on the peripheries of the Twitter network, they attacked one side (the Independentists) more than the other, inundated influential users on both sides with violently negative and inflammatory content, and exacerbated social conflict online. The researchers operationalized their research by first building the interaction network that unfolded on Twitter during the referendum and then scoring the tweets to measure their sentiment. They then built the semantic network of words that co-occurred in the tweets. Not only were tweets sent by bots to humans more negative; some of the negative hashtags that appeared in the semantic network of popular hashtags were almost always injected into the discourse by bots.

The Use of Social Media Data in Other Fields

The use of social media data is obviously not limited to the field of political communication. The techniques discussed and illustrated have also seen widespread use in other application domains as well.

In the area of business and marketing, for instance, social media data can be analyzed to detect changes in public opinion toward products and predict corporate crises (Bi, Zheng, & Liu, 2014) or to study the effectiveness of product placement on social media (Liu, Chou, & Liao, 2015). In crisis communication, the analysis of social media data has proven to be useful as well (Gill, Alam, & Eustace, 2014). Sentiment analysis of tweets can be used to monitor the spread of diseases (Ji, Chun, Wei, & Geller, 2015). More generally, in the area of public health, studies have found that Twitter data offer useful public health information (Paul & Dredze, 2011) and helps estimate county level health statistics (Culotta, 2014). Several studies have also found significant correlations between the volume of tweets and flu rates in the United States (Lampos & Cristianini, 2010). Likewise, sentiment analysis of social media data has been used in the context of mental health research to develop statistical measures that help discover shifts in suicidal tendencies (De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016).

All these studies use similar approaches to mine opinion and sentiments from publicly shared content in social media and analyze temporal trends in volume of content to determine whether and how they correlate with other statistical indicators. As with every type of empirical research, the value of the measures that can be extracted from social media activity depends on their ability to explain social phenomena, which (as the example discussed so far suggests) can range significantly, from political behavior at the level of individuals or groups, to trends in health behavior at the societal level.

Challenges in the Quantitative Analysis of Social Media

The phenomenal growth of the use of social media data for research is not just a testament to how this approach advances our understanding of the social world; it is also indicative of the relative ease of access to data that researchers now enjoy (see "Accessing Social Media Data" section). For all its benefits, the use of social media data can also often undermine the research process if appropriate measures are not taken to protect users' privacy and guarantee that basic ethical principles are followed. The next two sections consider some of the limitations associated with the quantitative analysis of social media data and ways to ensure an ethical approach to its analysis.

Limitations of Social Media Data

One of the primary things to consider when using social media data is the population to which the findings generalize. Twitter provides a great example of what is known as the *streetlight effect*. The streetlight effect refers to the phenomenon that researchers study something only because it is available, irrespective of how useful it is. The name comes from the hypothetical story of a person who, after having lost his key in a dark alley searches for it only under the streetlight—because that is where the light is. The main reason behind Twitter's continued use in public opinion and political communication research stems from this idea: the ease of access to its data. However, as studies have shown, Twitter users are not perfectly representative of the population of interest, namely the voting public (Barberá & Rivero, 2015). Similarly, different social media platforms attract different kinds of people, from various different demographics (Hargittai, 2015), and studies using data from different platforms can be limited in their generalizability. However, while, "[n]on representative data are bad for out-of-sample generalizations", they can be "quite useful for within-sample comparisons" (Salganik, 2017, p. 29). Thus, instead of dismissing observational studies using nonrepresentative found data as "limited," efforts should be made to try and do more observational studies using different data sets, each of which is representative of a different sample. "Estimates from many different groups, will do more to advance social research, than a single estimate from a probabilistic random sample" (Salganik, 2017, p. 33).

The second limitation of relying on social media data is the non-neutrality of these data sets. The "big-ness" of digital data, for example, is not necessarily a guarantee of their appropriateness for a research question.

Worse, it can often amplify preexisting biases. In other words, the bigger the data set, the higher the chances of a *systemic error* in results obtained using such a data set (McFarland & McFarland, 2015). An example of a systemic bias on social media is closely related to its public "social" nature. In other words, people share only a *particular type* of content on these platforms, while withholding other types owing to a variety of psychological and social factors like social desirability dynamics. Big data also have biases by design. The manner in which these data are collected is largely unknown to researchers and academics who use these data. Digital trace data can be more a reflection of the business interests of the companies that collect these data than a reflection of actual human activity (Tufekci, 2014). The algorithms that are used for the purposes of collecting these data are "black boxes," unknown even to the engineers who design them (Pasquale, 2015). A related concern is how these algorithms can potentially be confounding the results. Long-term studies can also suffer depending on how these methods of collection change. These changes can happen due to unforeseen circumstances (e.g., Facebook blocking access to their APIs—in response to claims about its use by Russian actors to influence the U.S. election) or a change in design (e.g., Twitter doubling its character limit). Longitudinal studies using social media data sources that could have potentially changed during the period being analyzed, should be treated with caution, and insights gleaned from them should be subject to such considerations as well.

Dean Freelon (2018) makes the more general but important point regarding the issue of access to social media data: because it depends entirely on the whims of the corporations that own these platforms (as Facebook suddenly blocking access to their APIs demonstrates), exclusively relying on formal APIs for conducting social media research can be debilitating, short-sighted, and even dangerous. He therefore recommends that researchers using social media have alternative ways to access data like web scraping and be cognizant of the terms and conditions of using the APIs that they do use.

Privacy and Ethics

The rise of big data has also led to the corresponding rise of "big sensitive data"—for instance, digitized health records and financial data. Such data sets can be easily misused by malicious actors, which can then be used to harm individuals (Timberg, Dwoskin, & Fung, 2017). Misuse of big data, however, is not limited to "sensitive information" only. Even anonymized and apparently innocuous data sets (like those obtained using Facebook or Twitter APIs), when linked with other relevant data sets, like survey responses—can be effectively deanonymized, as computer scientists have shown (Domingo-Ferrer et al., 2013). The naive use of data can also have unintended consequences for certain individuals, as was witnessed in two different situations: the first, when researchers made Facebook profiles of university goers publicly available (Zimmer, 2010) and again, during the WikiLeaks data dump from Saudi Arabia in 2016. It is also important to remember that no observational data set is neutral. They capture and reflect several biases and prejudices that are prevalent in society (Boyd & Crawford, 2012; see also Gonzaález-Bailoón, 2017; Pasquale, 2015; Tufekci, 2014). Using these data sets and making predictions or gleaning insights from them without considering their biases, can only propagate these biases and strengthen such prejudices, which in turn could be dangerous in the long run.

Similar concerns plague experiments using social media as well. In this regard, the Facebook emotion experiment that was briefly mentioned earlier (Kramer et al., 2014) comes to mind. In this study (an online field experiment), researchers at Facebook artificially manipulated the news feeds of Facebook users by adding and removing positive and negatively valanced posts from their friends in their Facebook feed. They observed that depending on which users saw more or less “happier” posts, their Facebook activity became more or less “happy.” What was disconcerting about this experiment was that not just that the results were statistically significant, but that the participants had no knowledge that this was happening to them. Even though the authors issued an apology later, and the journal published an expression of concern, this study highlights the potential pitfalls of the same affordances of digital technology that are otherwise celebrated.

Conclusion

This entry has discussed a large number of studies that use social media data in conjunction with quantitative methods. In this section, we summarize some of the key methodological approaches used in this domain. These methodologies involve tools for data collection and tools for analysis. Tools for data collection include queries through APIs, web scraping, merging survey data with social media activity, online experiments, and interventions with bots. The tools employed for analysis include natural language processing, network science techniques, machine learning, and computer vision.

We have discussed how social media data can be used for conducting descriptive studies or for running field experiments. Descriptive studies are useful to describe quantitatively social phenomena or to compare social dynamics as they emerge in different contexts. Field experiments are useful for answering cause–effect questions in a setting that is potentially more realistic and externally valid than the settings created to run lab experiments. The first step in many of the descriptive studies is collecting digital trace data. This can be achieved by means of formal APIs, by the use of third party tools that indirectly use formal APIs, or by techniques such as scraping. In field experiments on social media, researchers are able to gather additional information from social media users by surveying them and asking them questions about beliefs or behavior difficult to quantify through other means. Moreover, trace data can be artificially generated in field experiments by making subjects interact with bots (i.e., automated programs masquerading as real social media users).

Once the data are gathered, a range of computational techniques can then be used for analysis. These methods offer tools to analyze networks, transform text into indicators of public opinion, or to address prediction problems through the application of machine learning and topic modeling. Often, multiple methods are used in tandem.

Many of the current tools for data collection are likely to change as technologies evolve (e.g., the policies to access data through APIs change over time). However, the key elements of research design and the methodologies discussed in this entry will remain core pillars in the cumulative research agenda addressing the impact that social media activity has on social life, in all its ramifications (e.g., political behavior, health campaigns, social mobilization, viral marketing).

The use of social media data for social science might be a relatively recent phenomenon (compared to other data sources that have long been used in social science research), but it is safe to assume that this field of academic inquiry will only grow in the future. This growth will be driven by the increasing importance of social media in our lives, and the increasingly innovative ways in which researchers will process and analyze digital trails. Attempting to build a single toolbox for the quantitative analysis of social media data is a futile exercise since social media platforms and the nature of the data they generate will change over time. Very specific methods that are widely used today will become obsolete in the future, and the newer kinds of data that will then become available for access will require newer tool kits and likely more sophisticated methods for analysis.

While social media platforms will change, research with social media data will continue to be either observational or experimental, and new ways of integrating those research designs with survey data will surely keep on emerging. Likewise, the focus of analysis will also be either the structure of interactions or the content of the information being exchanged (or both). These are the general research parameters that will not change in the quantitative study of social media. The examples offered in this entry illustrate specific domains in which these parameters have been made operational. Social media platforms will surely morph in how they operate or give access to data, but solid research design will still rely on the choices and possibilities discussed here.

Further Readings

Gonzaález-Bailoón, S. (2017). *Decoding the social world: Data science and the unintended consequences of communication*. Cambridge, MA: MIT Press.

Ignatow, G., & **Mihalcea, R.** (2016). *Text mining: A guidebook for the social sciences*. Los Angeles, CA: SAGE.

Mejova, Y., **Weber, I.**, & **Macy, M. W.** (Eds.). (2015). *Twitter: A digital socioscope*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781316182635

Russell, M. A. (2013). *Mining the social web*. Stebastopol, CA: O'Reilly Media.

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.

Watts, D. J. (2011). *Everything is obvious: Once you know the answer*. New York, NY: Crown Business.

References

An, J., **Quercia, D.**, & **Crowcroft, J.** (2013). Fragmented social media. In *Proceedings of the 22nd International Conference on World Wide Web–WWW '13 Companion* (pp. 51–52). New York, NY: ACM Press. doi:10.1145/2487788.2487807

Aral, S., & **Walker, D.** (2012). Identifying influential and susceptible members of social networks. *Science*, 337, 337–341. doi:10.1126/science.1215842

Bail, C., **Argyle, L.**, **Brown, T.**, **Bumpus, J.**, **Chen, H.**, **Hunzaker, M. B.** … **Volfovsky, A.** (2018). Exposure to opposing views can increase political polarization: Evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences*, 115, 1–6. doi:10.17605/OSF.IO/4YGUX

Bakshy, E., **Messing, S.**, & **Adamic, L. A.** (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348, 1130–1132. doi:10.1126/science.aaa1160

Bakshy, E., **Hofman, J. M.**, **Watts, D. J.**, & **Mason, W. A.** (2011). Everyone's an influencer: Quantifying influence on Twitter categories and subject descriptors. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 65–74. doi:10.1145/1935826.1935845

Barbera, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23, 76–91. doi:10.1093/pan/mpu011

Barberá, P., **Jost, J. T.**, **Nagler, J.**, **Tucker, J. A.**, & **Bonneau, R.** (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26, 1531–1542. doi:10.1177/0956797615594620

Barberá, P., & **Rivero, G.** (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33, 712–729. doi:10.1177/0894439314558836

Barnett, G. A., & **Benefield, G. A.** (2017). Predicting international Facebook ties through cultural homophily and other factors. *New Media and Society*, 19, 217–239. doi:10.1177/1461444815604421

Bi, G., **Zheng, B.**, & **Liu, H.** (2014). Secondary crisis communication on social media: The role of corporate response and social influence in product-harm crisis. In *PACIS 2014 Proceedings*. Retrieved from http://aisel.aisnet.org/pacis2014/93

Blei, D. M., **Ng, A. Y.**, & **Jordan, M. I.** (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*. doi:10.1162/jmlr.2003.3.4-5.993

Bode, L., **Vraga, E. K.**, **Borah, P.**, & **Shah, D. V.** (2014). A new space for political behavior: Political social networking and its democratic consequences. *Journal of Computer-Mediated Communication*. doi:10.1111/jcc4.12048

Bond, R. M., **Fariss, C. J.**, **Jones, J. J.**, **Kramer, A. D. I.**, **Marlow, C.**, **Settle, J. E.**, & **Fowler, J. H.** (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*. doi:10.1038/nature11421

Bonilla, Y., & **Rosa, J.** (2015). #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist*. doi:10.1111/amet.12112

Boulianne, S. (2016). Online news, civic awareness, and engagement in civic and political life. *New Media & Society*, 18, 1840–1856. doi:10.1177/1461444815616222

Boyd, D., & **Crawford, K.** (2012). Critical questions for big data. *Information, Communication & Society*, 15, 662–679. doi:10.1080/1369118X.2012.678878

Casas, A., & **Williams, N. W.** (2018). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*. doi:10.1177/1065912918786805

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. doi:10.1111/jcom.12084

Conover, M. D., **Gonçalves, B.**, **Ratkiewicz, J.**, **Flammini, A.**, & **Menczer, F.** (2011). Predicting the political alignment of Twitter users. In *Proceedings—2011 IEEE international conference on privacy, security, risk and trust and IEEE International conference on social computing, PASSAT/SocialCom 2011*. doi:10.1109/PASSAT/SocialCom.2011.34

Conover, M., **Ratkiewicz, J.**, & **Francisco, M.** (2011). Political polarization on twitter. In *ICWSM 2011 International Conference on Weblogs and Social Media*. doi:10.1021/ja202932e

Culotta, A. (2014). Estimating county health statistics with twitter. doi:10.1145/2556288.2557139

De Choudhury, M. (2011). Tie formation on Twitter: Homophily and structure of egocentric networks. *Proceedings—2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 465–470. doi:10.1109/PASSAT/SocialCom.2011.177

De Choudhury, M., **Kiciman, E.**, **Dredze, M.**, **Coppersmith, G.**, & **Kumar, M.** (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI Conference*, 2016, 2098–2110. doi:10.1145/2858036.2858207

Del Vicario, M., **Zollo, F.**, **Caldarelli, G.**, **Scala, A.**, & **Quattrociocchi, W.** (2017). Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, 50, 6–16. doi:10.1016/j.socnet.2017.02.002

Dodds, P. S., & **Danforth, C. M.** (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11, 441–456. doi:10.1007/s10902-009-9150-9

Domingo-Ferrer, J., **Torra, V.**, **El Emam, K.**, **Dankar, F. K.**, **Issa, R.**, **Jonker, E.**, … Bash, E. (2013). How to break anonymity of the Netflix prize dataset. *Journal of the American Medical Informatics Association: JAMIA*, 8, 24. doi:10.1017/CBO9781107415324.004

Freelon, D. (2018). Computational research in the post-Api age. *Political Communication*, 35, 665–668. doi:10.1080/10584609.2018.1477506

Gil de Zúñiga, H., **Jung, N.**, & **Valenzuela, S.** (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*. doi:10.1111/j.1083-6101.2012.01574.x

Gill, A., **Alam, S.**, & **Eustace, J.** (2014). Using social architecture to analyzing online social network use in emergency management. *AMCIS 2014 Proceedings*, (February 2016), 1–12. Retrieved from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1060&context=amcis2014

Goel, S., **Anderson, A.**, **Hofman, J.**, & **Watts, D.** (2015). The structural virality of online diffusion. *Management Science*, 1909, 1–32. doi:10.1145/2229012.2229058

Gonzaález-Bailoón, S. (2017). *Decoding the social world: Data science and the unintended consequences of communication*. MIT Press. Retrieved from https://mitpress.mit.edu/books/decoding-social-world

González-Bailón, S., **Borge-Holthoefer, J.**, **Rivero, A.**, & **Moreno, Y.** (2011). The dynamics of protest recruitment through an online network. *Scientific Reports*, 1, 1–7. doi:10.1038/srep00197

González-Bailón, S., & **Wang, N.** (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks*, 44. doi:10.1016/j.socnet.2015.07.003

Grinberg, N., **Joseph, K.**, **Friedland, L.**, **Swire-Thompson, B.**, **Grinberg, N.**, & **Lazer, D.** (2019). Fake news on Twitter during the 2016 U.S. Presidential election. *Science*, 363, 374–378. doi:10.1126/science.aau2706

Guess, A., **Lyons, B.**, **Nyhan, B.**, & **Reifler, J.** (2018). Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think. Retrieved from https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf

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, 63–76. doi:10.1177/0002716215570866

Jackson, S. J., & **Welles, B. F.** (2015, August). #Ferguson is everywhere: Initiators in emerging counterpublic networks. *Information, Communication and Society*, 4462, 1–22. doi:10.1080/1369118X.2015.1106571

Jaidka, K., **Zhou, A. Y.**, & **Lelkes, Y.** (2018). Brevity is the soul of Twitter: The constraint affordance and political discussion. *SSRN*, 9, 34–40. doi:10.2139/ssrn.3287552

Ji, X., **Chun, S. A.**, **Wei, Z.**, & **Geller, J.** (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, 5, 13. doi:10.1007/s13278-015-0253-5

King, G., **Pan, J.**, & **Roberts, M. E.** (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, 345. doi:10.1126/science.1251722

King, G., **Schneer, B.**, & **White, A.** (2017, November). How the news media activate public expression and influence national agendas. *Science*, 780, 776–780.

Kramer, A. D. I., **Guillory, J. E.**, & **Hancock, J. T.** (2014). Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111, 10779–10779. doi:10.1073/pnas.1412469111

Lampos, V., & **Cristianini, N.** (2010). Tracking the flu pandemic by monitoring the social web. In *2010 Second International Workshop on Cognitive Information Processing, CIP2010*. doi:10.1109/CIP.2010.5604088

Liu, S. H., **Chou, C. H.**, & **Liao, H. L.** (2015). An exploratory study of product placement in social media. *Internet Research*, 25, 300–316. doi:10.1108/IntR-12-2013-0267

McFarland, D. A., & **McFarland, H. R.** (2015). Big data and the danger of being precisely inaccurate. *Big Data & Society*, 2. doi:10.1177/2053951715602495

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39, 629–649. doi:10.1007/s11109-016-9373-5

O'Connor, B., **Balasubramanyan, R.**, **Routledge, B. R.**, & **Smith, N. A.** (2010). From tweets to polls: Linking text sentiment to public opinion time series Brendan. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. https://doi.org/citeulike-article-id:7044833

Pasquale, F. (2015). *The black box society the secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Paul, M. J., & **Dredze, M.** (2011). You are what you tweet: Analyzing Twitter for public health. *Fifth International AAAI Conference on Weblogs and Social Media*, 265–272.

Peng, Y., & **Jemmott, J. B.** (2018). Feast for the eyes: Effects of food perceptions and computer vision features on food photo popularity. *International Journal of Communication*, 12, 313–336.

Pennacchiotti, M., & **Popescu, A.-M.** (2011). A machine learning approach to twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*. doi:10.1145/2542214.2542215

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.

Skoric, M. M., **Zhu, Q.**, **Goh, D.**, & **Pang, N.** (2016). Social media and citizen engagement: A meta-analytic review. *New Media and Society*. doi:10.1177/1461444815616221

Smith, A., & **Anderson, M.** (2018). *Social media use 2018*. Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

Stella, M., **Ferrara, E.**, & **De Domenico, M.** (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1803470115

Stieglitz, S., & **Dang-Xuan, L.** (2013). Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29, 217–248. doi:10.2753/MIS0742-1222290408

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton, NJ: Princeton University Press.

Theocharis, Y., **Lowe, W.**, **van Deth, J. W.**, & García-Albacete, G. (2015). Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information Communication and Society*. doi:10.1080/1369118X.2014.948035

Timberg, C., **Dwoskin, E.**, & **Fung, B.** (2017, September 7). Data of 143 million Americans exposed in hack of credit reporting agency Equifax. *The Washington Post*. Retrieved from https://www.washingtonpost.com/business/technology/equifax-hack-hits-credit-histories-of-up-to-143-million-americans/2017/09/07/a4ae6f82-941a-11e7-b9bc-b2f7903bab0d_story.html?utm_term=.21c876283c6d

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19. doi:10.5210/fm.v19i7.4901

Ugander, J., **Backstrom, L.**, **Marlow, C.**, & **Kleinberg, J.** (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109, 5962–5966. doi:10.1073/pnas.1116502109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1116502109

Ugander, J., **Karrer, B.**, **Backstrom, L.**, & **Marlow, C.** (2013). The anatomy of the Facebook social graph. *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation—RepSys*, 13, 15–22. doi:10.1145/2532508.2532512

Vosoughi, S., **Roy, D.**, & **Aral, S.** (2018). The spread of true and false news online. *Science*, 359, 1146–1151. doi:10.1126/science.aap9559

Watts, D. J. (2004). *Six degrees: The science of a connected age*. Retrieved from https://books.google.com/books/about/Six_Degrees_The_Science_of_a_Connected_A.html?id=1gueFWR7qjoC

Welles, B. F., & **González-Bailón, S.** (Eds.). (2018). *The Oxford handbook of networked communication*. Oxford University Press. doi:10.1093/oxfordhb/9780190460518.001.0001

Zamal, F. Al, **Liu, W.**, & **Ruths, D.** (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. doi:10.9770/jesi.2015.3.1(2)T

Zhao, W. X., **Jiang, J.**, **Weng, J.**, **He, J.**, **Lim, E.-P.**, **Yan, H.** … Xiaoming. (2011). Comparing Twitter and traditional media using topic models. In **P. Clough**, **C. Foley**, **C. Gurrin**, **G. J. F. Jones**, **W. Kraaig**, **H. Lee**, & **V. Mudoch** (Eds.), *Advances in information retrieval. ECIR 2011. Lecture notes in computer science* (Vol. 6611, pp. 339–349). doi:10.1007/978-3-642-20161-5_34

Zimmer, M. (2010). But the data is already public: On the ethics of research in Facebook. *Ethics and Information Technology*, 12, 313–325. doi:10.1007/s10676-010-9227-5