

The good, the bad, and the ugly: uncovering novel research opportunities in social media mining

Huan Liu¹  · Fred Morstatter¹ · Jiliang Tang² · Reza Zafarani³

Received: 3 August 2016 / Accepted: 22 August 2016 / Published online: 2 September 2016
© Springer International Publishing Switzerland 2016

Abstract Big data is ubiquitous and can only become bigger, which challenges traditional data mining and machine learning methods. Social media is a new source of data that is significantly different from conventional ones. Social media data are mostly user-generated, and are big, linked, and heterogeneous. We present the good, the bad and the ugly associated with the multi-faceted social media data and exemplify the importance of some original problems with real-world examples. We discuss bias in social media data, evaluation dilemma, data reduction, inferring invisible information, and big-data paradox. We illuminate new opportunities of developing novel algorithms and tools for data science. In our endeavor of employing the good to tame the bad with the help of the ugly, we deepen the understanding of ever growing and continuously evolving data and create innovative solutions with interdisciplinary and collaborative research of data science.

Keywords Social media · Data mining · Evaluation · Big-data paradox · Data analytics

1 Introduction

Social media is an unprecedented means of mass communication. It differs from traditional media such as TV channels, radio stations, movie theaters, news papers, classrooms, snail mail, telephones, fax machines, etc. These traditional media communicate largely in two modes: *1-to-many* or *1-to-1*. Social media allows for a new mode of *many-to-many* communications anytime anywhere [43]. The rise of social media opens the door for many new phenomena. With easy-to-use user interface, the prevalence of mobile devices, and a subsequent disappearing communications barrier, everyone can be a media outlet or content producer. Social media has its distinct characteristics. It enables rich user interactions via multi-modal connections and various types of relations among users and between users and other entities, resulting in plethora amounts of linked data, and produces user-generated content that massive, dynamic, extensive, instant, and extremely noisy. Social media is also a natural collaborative environment where crowdsourcing is made easy such that groups of special interests are formed for a wide range of purposes. Researchers repeatedly discover that social media networks follow some kind of power law distribution, exhibiting a long-tail phenomenon with many small groups. While all seems free, attention becomes precious and can be turned into monetary values or political forces. Social media is now an undisputed important source of rich data and if used appropriately, a new lens for many new types of study, a.k.a., computational social science or social computing.

Social media data can be obtained from publicly available sources via various means such as scraping, using

✉ Huan Liu
huan.liu@asu.edu

✉ Fred Morstatter
fred.morstatter@asu.edu

Jiliang Tang
jiliang.tang@cse.msu.edu

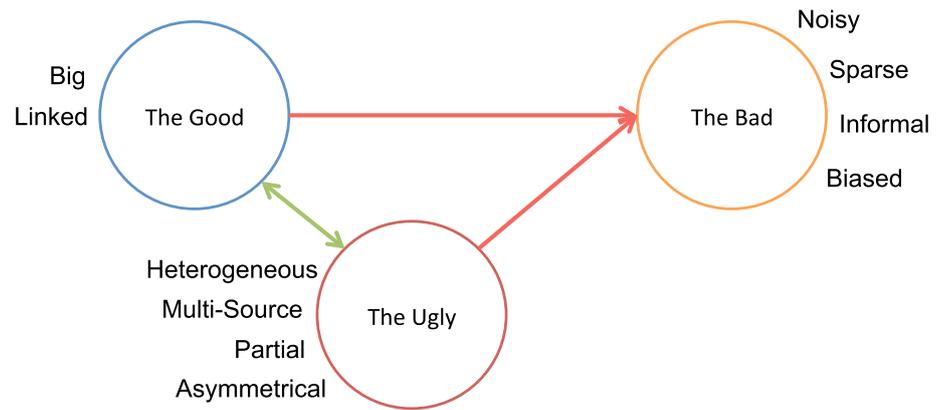
Reza Zafarani
rzafaran@syr.edu

¹ School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287-8809, USA

² Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

³ Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, USA

Fig. 1 An illustrative view of the relationship between the good, the bad, and the ugly of big social media data. The good and the ugly work together to overcome the challenges due to the bad



site-provided apps, and crawling. The possibility of obtaining social media data makes research on social media data feasible. In *Twitter Data Analytics* [18], for example, a begin-to-end process for Twitter data analytics is elaborated with four key steps: crawling, storing, analyzing, and visualizing. Social media data are just like conventional data, which is a potential treasure trove, but requires data mining to uncover hidden treasures. Social media mining faces novel challenges due to the distinct characteristics of social media data. We discuss the good, the bad and the ugly of social media data before presenting details on how we can seek out research opportunities from these unique challenges.

The good of social media data is that it is big and linked [22]. Here is an example of daily use of big and linked data. On a sunny Saturday early morning, two families were traveling from Phoenix to a high school south of Tucson for the 2016 Arizona State Mathcounts Competition, a large state-wide event for middle school students. According to a web search, it is a 2.5-h drive. The usual freeway Interstate 10 was slowed down and gradually stalled to a snail's pace just before Tucson due to a car crash early on, causing a large traffic jam. Two families took the trip independently in two vehicles. Family A followed the traffic, and without notice, became one member that made the long line longer. Family B took the alert from their mobile phone of slow traffic ahead and a detour advice, and got out of the freeway just in time. Family B drove on surface roads smoothly for an additional 3 miles. The difference is that Family A started their journey a bit earlier, but ended up at the destination more than half an hour later than Family B, barely before the math competition started. Both families used online maps to guide their trip. What made the difference? It is the big and linked data collected from the user patterns that are mined from collective mobile phone data.

The bad of social media data is that it is noisy, and data is often missing where it is most needed [9]. It is noisy because it is mostly user-generated and users can freely express themselves. Thus, gibberish or otherwise formal or informal writings, and incomplete sentences are often prevalent. The shortage of data is due to the power law distribution

of social networks, while nodes in the short head can be rich in data, nodes in the long tail have much fewer connections and much less information.

The ugly of social media data is that it is heterogeneous, of multiple sources, partial, and asymmetrical [41]. A user can have different connections with people, groups, or posts. A post can consist of text, URLs, audio, image, or video. A user can be active on different social media services, but only provides partial or minimal information on each service. More often than not, user–user or user–item relations are asymmetrical. For example, user A follows user B, but B may not follow back. User C likes item D, but D cannot like back.

Understanding the good, the bad and the ugly of social media data, we now discuss how we can tame the ugly and work with the good to overcome the bad and in the process, illustrate how we can find novel research opportunities for new discoveries. This process is illustrated in Fig. 1. In the following, we organize our discussion based on some novel challenges encountered in our research and use specific examples to demonstrate how innovative research can help approach these challenges.

2 Bias in social media

Social media data are an important part of linked social data. It affords a new look into human behavior at scales thought impossible only a few years ago. However, the observations and research findings made with social media data can only be generalized to human behavior if social media provides a representative description of human activity. In fact, recent studies have found evidence of many different sources of bias in social media data. This bias can come from demographic bias in social media. For example, the average age of Twitter users is much younger than the general population [26]. Malicious and automated accounts can produce massive amounts of content pollution, thus skewing the statistics of the site [31]. Also, the way that sites distribute their data can be biased [29], providing a skewed representation of their content through their user interfaces and APIs. In

another study, it was discovered that not only can the APIs be biased in the way they distribute data, but the mechanisms through which these data are distributed can be attacked [27]. This means that, in their current form, some APIs can be manipulated by bots, spammers, and other forms of content polluters in order to make their messages appear more salient to those using APIs to collect data.

Content polluters such as bots present a major obstacle to those studying social media data. Botnets can work in tandem to generate noise on social media, changing the statistics of the site. One example of this phenomenon occurred in early 2015, when a group of bots worked together to change the narrative surrounding a story.¹ In April 2015, a journalist who was publishing anti-Russia articles was killed. Twitter users began speculating that it could have been a Russian operative that killed him. Suddenly, a deluge of automated accounts came online and began tweeting the same identical message: “Ukrainians killed him...he was stealing one of their girlfriends”.² With so many bots tweeting this message, it quickly became one of the top posts about this incident. These bots were able to drown out the voice of the real people and pushed an agenda by the controller of the bots. This mass of identical tweets not only pushes these messages to the top of Twitter’s algorithms, but will also skew the statistics of the dataset, such as top terms, hashtags, and who the most important users are.

While content polluters are one issue that affects how well social media data represent human behavior, another is how the sites themselves share their data. Many social media sites provide access to the data produced on their sites through APIs. Researchers rely upon these APIs to collect data to verify their hypotheses. Therefore, it is essential that enough data be collected such that credible findings can be obtained. Previous research has, however, found evidence of bias in social media [29], e.g., the top hashtags of the tweets that come through Twitter’s APIs (i.e., 1%) can be significantly different from the top hashtags on all of Twitter during the same period. This has implications for research done on these APIs as it means that measurements taken from these API samples may not truly reflect what is unfolding on Twitter. Though 1% of Twitter data can still be very big, the finding suggests that care should be given by taking into account possible sample bias. The realization of some biases in social media data of many sorts suggests new research problems related to bot detection, neutralizing content pollution, estimating sampling bias and how to figure it out without ground truth, discovering additional biases in social media.

¹ <https://globalvoices.org/2015/04/02/analyzing-kremlin-twitter-bot/>.

² https://twitter.com/ASLuhn/status/571479498560028672/photo/1?ref_src=twsrc%5Etfw.

3 Needs for innovative evaluation methods

Evaluation is an indispensable component of social media mining [47] and serves an important role in data science. One type of common machine learning and data mining algorithms is classification [2]. One of the key steps in building a predictive model (a type of classification algorithms) in machine learning or data mining is to obtain ground truth. When we say ground truth data in this context, we mean a set of data that is tagged with correct labels so that we can better evaluate and verify if the predictive model can actually work as it is claimed so and different algorithms or models can be fairly compared in order to make solid advance in research and development. For example, to build a sentiment classifier [32] from social media posts, one must first curate a set of posts that are labeled with known sentiment. Although it is important to obtain quality ground truth, it is challenging to find ground truth for social media research. For the problem of bot detection [23], for instance, it is very difficult to know whether a user is truly a bot or not. If we would like to build a bot classifier, we have to learn it from data with both users and verified bots. While we may have some strong signals that tell us that a user is behaving erratically, without actually observing the user we cannot know for sure if that erratic behavior is truly because they are controlled by software.

This challenge invigorates novel approaches to confirm if they are likely to be a bot. One method is to recruit human annotators to manually inspect the users and to see if they are behaving like bots [7]. While this approach is usually accurate due to the skill of the human annotators, it is not scalable. To circumvent the issue of scalability from human annotation, automated approaches are also taken to labeling these users. For example, we can observe the way the site reacts to the users in order to see if they are a bot or not [42]. Most major social media platforms have inbuilt bot detection mechanisms that try to remove harmful users such as bots from their sites. In order to leverage this, we take two screenshots of the networks. First, we collect a sample of the network. Next, we wait for a set amount of time to allow the mechanisms in the social media site to work and delete the bot users. Finally, we re-crawl the social network and see which user accounts were suspended by the site. Those that are suspended are then marked as bots in our dataset, and those that are still active are marked as human. While this approach is very scalable due to its automatic nature, it has some key drawbacks. Social media sites try to avoid deleting users in order to keep a large user base, meaning that there are many false negatives. Also, it is possible to be suspended from most social networking sites for other reasons such as sharing copyright-infringing links, causing false positives. Another approach that seeks to be automated while also avoiding the issues that are inherent to using the

labels from the site are “honeypots”. Honeypots are automated social media accounts created by the researchers in order to lure other bots into following them. By learning the patterns that bots employ while operating on social networks, we can design honeypots exploit these patterns to attract bots to follow them [31]. Examples of these types of patterns include that bots tend to follow users who post on certain topics (e.g. hashtags, subreddits). On Twitter, bots tend to follow users who state that they “follow back” in their profile. By building honeypots and allowing them to operate on the social network for a set amount of time, we can collect a set of bots.

In natural language processing (NLP) research such as topic modeling, results are manually evaluated by researchers [11] due to the need for evaluating the degree of interpretability of topics, another task for which we do not have ground truth. A crowdsourcing approach is proposed by Chang et al. [6] to assign a quality score to the topics. In [30], Morstatter et al. propose a measure to assess how well humans can assign a “title” to topics. For example, in topics generated from newspaper text, how often does a topic assigned the definition “sports” actually primarily contain tokens from sports articles? Another way to view the semantic quality of topic modeling algorithms is to view the topics as clusters. When viewed from this way, there are two natural angles from which we could assess the topic clusters: the within-topic distance, and the between-topic distance. The measure proposed in [6] addresses the between-topic distance, and the measure proposed in [28] addresses the within-topic distance. This is done through a new measure which assesses how well the crowdsourced workers are able to single out words that do not belong to the topic.

A more challenging issue to evaluation is that more often than not, in social media research, it is inevitable that there is no ground truth at all. For example, in a study of migration between different social networking sites [19], on the one hand, it is obvious that the study would be unnecessary if we already knew there was some ground truth about migration or not; on the other hand, when the study presents some migration patterns, it is necessary to evaluate whether these patterns are haphazard or not. In other words, we still need to verify the findings when no ground truth is available. The reality in social media research forces us to seek evaluation techniques from different disciplines that can help scientifically verify research findings when no ground truth is available [47]. One application of this line of evaluation appears in topic modeling [3]. Topic modeling is the process of identifying topics from large bodies of text, a task that inherently lacks ground truth. Instead of seeing how well the predicted topics match some ground truth labels, it is suggested to measure how well the topics learned match a set of held-out documents. The measure they use to evaluate the match is called “perplexity”, and it works by measuring the

distribution of the topics against the distribution of the held-out documents [10]. In this way, they are able to measure the performance of topic modeling algorithms without the need to tag a single document. This is just one example of how we can estimate the performance of a predictor without ground truth.

4 Removing noise from social media data

Social media data are special in many ways with respect to conventional attribute-value data commonly used in classic data mining. Both types of data can be exceedingly large in terms of size and dimensionality. Social media data are linked via pervasively available social relations (i.e., friendships in Facebook), and mostly user-generated, thus is extraordinarily noisy. Social media data are typically high dimensional. For example, there are millions of terms in tweets, while high-quality images from Flickr could have millions of pixels. Usually only a small portion of features are relevant to a given social media mining task and others are irrelevant, redundant and noisy. Therefore, *its features are noisy* [37]. Users in social media can be both passive content consumers and active content producers and the quality of social media data is varied drastically from excellent content to abuse and spam. For example, more than 50 % of tweets are pointless babble and irritating spam. Two common characteristics of linked data such as concentrated linkage and autocorrelation can significantly reduce the effective size of instances for mining and learning [16]. Hence, *its data instances are noisy* [39]. Social media goes beyond the physical constraints of user relationship and allows one to be connected with many users of different relational types. For example, Twitter users have a small number of friends compared to the numbers of followers and followees they have [14]. Links with best friends, acquaintances and even spammers are usually mixed and given the sheer number of connections, it is difficult to differentiate them. Thus, *its links are noisy* [8].

As we learn from data mining 101 that “garbage in and garbage out” [13], it is essential to pre-process data for effective data mining. Therefore, it is intuitive and sensible to remove noisy features, noisy instances and noisy links before we proceed with social media mining. However, given the fact that we can often access to a small percentage of data (e.g., 1 % from Twitter), we ask what remains after noise removal. Following traditional data-preprocessing methods, it is very likely that little data remain given the large amounts of noise. They present unique challenges to noise removal. Novel research is required to take advantage of social media data with distinct properties such as links and multiple sources in addition to attribute-value data. For example, linked instances are more similar in terms of topics and feature distributions, while multiple sources could pro-

vide a more comprehensive view about social media data, making it possible to solve some problems unsolvable using a single source.

Linked and multi-view feature selection algorithms have recently been designed that take advantage of link information and multiple sources to select the most representative and informative features, respectively [36–38]. Both link information and multiple sources can provide additional, helpful constraints to advance the selection process—in the supervised scenario, these constraints can reduce the requirement of the amount of labeled data significantly, while they can help select better and more stable sets of features in the unsupervised setting. Among three types of noise, removing one type of noise can benefit removing others. For example, instances without noise can help the selection of useful features; and links without noise can provide better constraints to remove noisy instances. Therefore, it makes sense to develop joint-noise-removal frameworks to perform removing multiple types of noise in social media data simultaneously [39,40].

5 Discovering implicit negative links

Social networking sites make it easy for users to connect with, follow, or “like” each other. Most social networking sites therefore try to promote positive connections among users and help their growth in terms of users without mechanisms for negative encounters. This is a type of one-way connections that only allows for like; hence, it makes no distinction between indifference and dislike. As one’s social network grows, it is inevitable that users might not be benevolent toward each other, and implicit negative links could form. There is a solid rationale that few social networking sites allow online users to explicitly specify negative links [35]: (1) they are unwanted properties that could jeopardize the stability of online communities; (2) they could block information propagation; and (3) they could encourage escalating vengeance. However, recent work suggests that negative links have significant added value over positive links in various analytical tasks [21]. For example, a small number of negative links can significantly improve positive link prediction [12], and they can also improve the performance of recommender systems in social media [33]. Implicit negative links are not readily available, but they can be helpful in prediction tasks. This absence of explicit negative links presents a challenge. If we could mine implicit negative links, we could take advantage of negative links to advance social media applications.

A new challenge is how to find these implicit negative links, which galvanizes original research of discovering implicit negative links on social networking sites. Looking at the physical world, people can reveal their implicit disagreement, objection, or negative opinions in many indirect

ways, or signals characterized in [25]. Is it feasible for us infer the invisible negative links via signals indigenous to social media data. Recent studies reveal some interesting signals: (1) most triads in a signed network (or a network with both positive and negative links) satisfy balance and status theories [24]; (2) negative links present distinct properties in terms of clustering coefficient, reciprocity and transitivity from positive links [35]; (3) our foes are closer than random nodes to us in the positive network, typically within 2 or 3 hops, though our friends are the closest [34]; (3) there is a strong correlation between negative interactions and negative links, and the more negative interactions two users have, the more likely a negative link exists between them [34]; and (4) users with higher optimism (pessimism) are more likely to establish positive (negative) links than those with lower optimism (pessimism) [1]. These observations have laid the ground work to develop meaningful algorithms and further research on discovering the implicit negative links [34].

6 Big-data paradox

Social media data are undoubtedly big and offers countless opportunities for analyzing human behavior. Researchers of different disciplines welcome this new data source to study human behavior at scale. The question is whether social media is really big for our study. The unrelenting reality is that unfortunately, even with this seeming big data, the data at the individual level is often extremely limited for most users. Hence, we face a *big-data paradox*. We explain why this is a common situation in our research endeavor. On a social media site, many users are content consumers [5]. On Twitter, for example, more than 40% of the users have never tweeted.³ Another observation is that many users have few friends or connections. Both content generation and user degrees (connections) can be well described by the known *Pareto Principle* or the *80–20 Rule*. Basically, 80% of the content on a site is generated by 20% of the users [4]; and a few users have a huge number of followers or friends, and a large number of users have a few friends. Since we often focus our study of the users in the long tail of the distribution and they are the majority, it is certain that these users themselves contain or produce limited data at the individual level, thus, denoted as *thin data*.

This *big-data paradox* presents a unique challenge to social media mining: how can we conduct data analysis such as behavior analysis when we are inundated with large collections of thin data? Although this phenomenon happens on many social media sites, user data is not limited to a single site. Users often join multiple social media sites for various reasons. On each site, they leave barely minimum data, i.e.,

³ <http://goo.gl/2Xr9X>.

sparse or thin data. To study social media user behavior, we need as much data as possible. An intuitive solution is to automatically find a user's different profiles on the sites he registered on and concatenate his many sparse data to make thin data thicker. It is not an easy task [15, 17, 44, 45]. One has to consider two important constraints. First, to be able to analyze behavior of all users, one has to be able to study them with the amount of information that is guaranteed to be available for each and every one of the users. In other words, we have to be able to study users with the *minimum information* that is always available for any user. Second, one has to be able to accumulate data that belongs to the same user across sites. Hence, the same users should be identified across sites, however, with minimum information. This big-data paradox stimulated a new research problem of making thin data thicker in behavior analysis. Another associated challenge is how to use limited information to achieve our goals of social media mining.

There have been efforts to analyze user behavior with limited information. These studies have efficiently gleaned traces of human behavior in the information that is available for each individual. Because user data is sparse and spread across multiple sites, these methods are constrained to utilize the minimum information available in social media. Some applications successfully use these methods (1) to analyzed various user behaviors, such as migrations on social networking sites [20] or malicious activities [46] and (2) to identify the same users across sites with high accuracy [48].

7 Looking ahead

We start our discussion on the good, the bad, the ugly of social media, next employ some evident challenges to illuminate what the problems are, how difficult they are, and what new research opportunities these unprecedented challenges call for. Our elaboration is centered around five challenges: various types of biases in social media, different needs for evaluation, removing noise from social media data, discovering implicit negative links, and big-data paradox. With specific examples in our research, we make an attempt to convey what we believe in—novel and interesting research problems can be found where challenges are. Social media opens the door for people of all walks of life and, in the meantime, offers an unparalleled platform for researchers and scientists to study human behavior and activities at scale. The five challenges and their discussions are far from being completing, but only serve as a tangible means to help illuminate the vast potential and unlimited possibility of new challenges and original research. In order to facilitate collaboration and peer evaluation, we maintain two data and/or

algorithms repositories: the social computing repository⁴) and the feature selection repository (scikit-feature Python repository⁵). With illustrative work presented in this article, we are hopeful and look forward to steady advancement of research and development in social media mining in particular, and in data science in general.

Acknowledgments This material is based upon works supported by, or in part by, the U. S. Army Research Office under Grant Number #025071, National Science Foundation under Grant Number (IIS-1217466) and Office of Naval Research under Grant Numbers (N000141410095, N00014-16-1-2257). The authors are grateful to the former and current members of ASU DMML laboratory and collaborators in these projects.

References

1. Beigi, G., Tang, J., Liu, H.: Signed link analysis in social media networks. In: Tenth International AAAI Conference on Web and Social Media (2016)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2001)
3. Blei, D.M.: Probabilistic topic models. *Commun. ACM* **55**(4), 77–84 (2012)
4. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 1–14. ACM (2007)
5. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Trans. Netw. (TON)* **17**(5), 1357–1370 (2009)
6. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. *Adv. Neural Inf. Process. Syst.* 288–296 (2009)
7. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 21–30. ACM (2010)
8. Gao, H., Wang, X., Tang, J., Liu, H.: Network denoising in social media. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 564–571. ACM (2013)
9. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 211–220. ACM (2009)
10. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proc. Natl. Acad. Sci.* **101**(suppl 1), 5228–5235 (2004)
11. Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21** (3), 267–297 (2013)
12. Guha, R., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of trust and distrust. In: Proceedings of the 13th international Conference on World Wide Web, pp. 403–412. ACM (2004)
13. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)

⁴ <http://socialcomputing.asu.edu/>.

⁵ <http://featureselection.asu.edu/scikit-feature/>.

14. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. In: Available at SSRN 1313405 (2008)
15. Iofciu, T., Fankhauser, P., Abel, F., Bischoff, K.: Identifying users across social tagging systems. In: ICWSM, pp. 522–525 (2011)
16. Jensen, D., Neville, J.: Linkage and autocorrelation cause feature selection bias in relational learning. *ICML* **2**, 259–266 (2002)
17. Korula, N., Lattanzi, S.: An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.* **7**(5), 377–388 (2014)
18. Kumar, S., Morstatter, F., Liu, H.: *Twitter Data Analytics*. Springer, New York (2014)
19. Kumar, S., Zafarani, R., Liu, H.: Understanding user migration patterns in social media. In: AAAI (2011)
20. Kumar, S., Zafarani, R., Liu, H.: Understanding user migration patterns in social media. In: Twenty-Fifth AAAI Conference on Artificial Intelligence (2011)
21. Kunegis, J., Preusse, J., Schwagereit, F.: What is the added value of negative links in online social networks? In: Proceedings of the 22nd International Conference on World Wide Web, pp. 727–736. ACM (2013)
22. Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Life in the network: the coming age of computational social science. *Science* **323**(5915), 721 (2009)
23. Lee, K., Eoff, B.D., Caverlee, J.: Seven months with the devils: a long-term study of content polluters on twitter. In: ICWSM, pp. 185–192 (2011)
24. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World wide web, pp. 641–650. ACM (2010)
25. Loudon, A.: Know your enemy. *ANNALS Am. Acad. Polit. Soc. Sci.* **222**, 26–31 (1942)
26. Mislove, A., Lehmann, N., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the demographics of twitter users. *ICWSM* **11**, 5 (2011)
27. Morstatter, F., Dani, H., Sampson, J., Liu, H.: Can one tamper with the sample api? Toward neutralizing bias from spam and bot content. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 81–82. International World Wide Web Conferences Steering Committee (2016)
28. Morstatter, F., Liu, H.: A novel measure for coherence in statistical topic models. In: Association for Computational Linguistics, pp. 543–548 (2016)
29. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. In: ICWSM (2013)
30. Morstatter, F., Pfeffer, J., Mayer, K., Liu, H.: Text, topics, and turkers: A consensus measure for statistical topics. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 123–131. ACM (2015)
31. Morstatter, F., Wu, L., Nazer, T.H., Carley, K.M., Liu, H.: A new approach to bot detection: striking the balance between precision and recall, pp. 1–8 (2016)
32. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2**(1–2), 1–135 (2008)
33. Tang, J., Aggarwal, C., Liu, H.: Recommendations in signed social networks. In: Proceedings of the 25th International Conference on World Wide Web, pp. 31–40. International World Wide Web Conferences Steering Committee (2016)
34. Tang, J., Chang, S., Aggarwal, C., Liu, H.: Negative link prediction in social media. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 87–96. ACM (2015)
35. Tang, J., Chang, Y., Aggarwal, C., Liu, H.: A survey of signed network mining in social media. *ACM Comput. Surv.* **42**, 1–37 (2016)
36. Tang, J., Hu, X., Gao, H., Liu, H.: Unsupervised feature selection for multi-view data in social media. In: SDM, pp. 270–278. SIAM (2013)
37. Tang, J., Liu, H.: Feature selection with linked data in social media. In: SDM, pp. 118–128. SIAM (2012)
38. Tang, J., Liu, H.: Unsupervised feature selection for linked social media data. In: KDD, pp. 904–912. ACM (2012)
39. Tang, J., Liu, H.: Coselect: feature selection with instance selection for social media data. In: SDM, pp. 695–703. SIAM (2013)
40. Tang, J., Liu, H.: Feature selection for social media data. *ACM Trans. Knowl. Discov. Data (TKDD)* **8**(4), 19 (2014)
41. Tang, L., Liu, H.: Community detection and mining in social media. *Synth. Lect. Data Min. Knowl. Discov.* **2**(1), 1–137 (2010)
42. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 243–258. ACM (2011)
43. Zafarani, R., Abbasi, M.A., Liu, H.: *Social Media Mining: An Introduction*. Cambridge University Press, Cambridge (2014)
44. Zafarani, R., Liu, H.: Connecting corresponding identities across communities. *ICWSM* **9**, 354–357 (2009)
45. Zafarani, R., Liu, H.: Connecting users across social media sites: a behavioral-modeling approach. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 41–49. ACM (2013)
46. Zafarani, R., Liu, H.: 10 bits of surprise: Detecting malicious users with minimum information. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 423–431. ACM (2015)
47. Zafarani, R., Liu, H.: Evaluation without ground truth in social media research. *Commun. ACM* **58**(6), 54–60 (2015)
48. Zafarani, R., Tang, L., Liu, H.: User identification across social media. *ACM Trans. Knowl. Discov. Data (TKDD)* **10**(2), 16 (2015)