What Social Media Data Should I Use in My Research?: A Comparative Analysis of Twitter, YouTube, Reddit, and the New York Times Comments

Dongho Choi, Ziad Matni, Chirag Shah
School of Communication & Information (SC&I)
Rutgers, The State University of New Jersey
4 Huntinton St, New Brunswick, NJ 08901, USA
{dongho.j.choi,zmatni,chirags}@rutgers.edu

ABSTRACT

The emergence of popular social media allows people to generate and share information in unprecedented and multiple ways. While studies concerning social media data are plentiful and can be found across a variety of scholarly disciplines, there has been little research done that compares multiple social media while considering topics and goals of an investigation. We collected and analyzed comparable data from four different social media sources: *Twitter, YouTube, Reddit,* and the *New York Times* website, across five research topics. We then analyzed these data sets across several criteria, such as richness and the diversity of information that they contain. Our analysis shows the trade-offs between different social media services and also provides a new framework to compare cross-platform data collection and analysis utility.

Author Keywords

Social Media, Data Collection, Research Goal(s),
Comparative Analysis

INTRODUCTION

After the emergence of Web 2.0, a rich amount of information has been generated and shared by ordinary users of the Internet. This information conveys all manner of news, facts, opinion, emotions, entertainment, and multiple creative outlets, among other things. Moreover, as numerous forms of social media have emerged, multiple ways in which people generate and share information have also emerged. The manner in which people create information and content, and share it through their social media networks varies across the particular social media they use.

A large number of scholarly works have used social media as sources of data that reveals the information behavior of its users. However, there has been little research that compares multiple social media while considering research topics and goals. As more diverse platforms of social media emerge, and as more researchers and scholars use social media for their data collection, the choice of appropriate social media targets for specific research topics and goals has gained importance. In this paper, we address this by collecting and analyzing data from different social media sources. Specifically, we collected comparable data sets on five topics from *Twitter, YouTube, Reddit,* and the *New York Times* website. We then analyzed these data sets across several criteria, such as richness and the diversity of information they contain. The analyses, on the one hand shows the trade-offs between different social media services, and on the other hand provides a new framework to compare cross-platform data collection and analysis utility.

RELATED WORK

Hughes et al. (2012) examined how people use Twitter and Facebook as tools for social interaction and information exchange, and found that the personality of the user correlates with online socializing, information exchange, and one's preference of using a particular social media over another. Since social media acts as an active and effective channel for opinion transmission (Zhao et al., 2011), scholars in politics have investigated political behaviors and patterns on social media in which people produced and consumed political information (Loader and Mercea, 2011; Shirky, 2011) Wallsten (2010) used viral video data taken from social media to explore the factors that lead these videos to spread and penetrate the political discourse.

Social media also has been considered to be fast-developing and effective ways for word-of-mouth (WOM) effect. WOM effect encompasses informal communication between consumers about different products and services (Liu, 2006). Opinions and suggestions related to products and services are generated and shared on several platforms of social media (Bakshy et al., 2011; Jansen et al., 2009).

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark

Scholars have studied social media with both exploratory and predictive modeling goals in mind. For example, Hawn (2009) examined social media in order to understand how people dealt with particular topics in various domains, finding in health care practices social media are reshaping the way doctors and patients interact. Predictive models have been generated from social media data, for instance, the rate at which tweets are posted about particular movies has been used to predict their box office revenue (Asur and Huberman, 2010) as well as stock market (Gilbert and Karahalios, 2010).

Yang et al. (2015) compared social media with regard to as tools to address questions, finding the innate characteristics of major social media for the design of social Q&A services. Users in different social media present distinct behaviors regarding information interaction. Benevenuto et al. (2009) analyzed user workloads in social networks to find key features that differentiate social actions, such as how frequently people connect each other and for how long, as well as the types and sequences of user activities on these platforms.

RESEARCH QUESTIONS

While there are plenty of studies that collect and analyze data from a social media source, there has been little research conducted on a comparative analysis of multiple platforms of social media, from the perspective of information sources for social research. From the literature review, we found that (1) peoples' engagement in social media, including direct and indirect activities, can be observed in very diverse platforms; (2) a large portion of previous research has gathered information from social media for exploratory studies or for generating predictive models; and (3) depending on research topics and questions, different social media have been used, showing different characteristics, advantages, and disadvantages. Considering previous explorations we address some questions about working with social media:

1. How do different social media reveal different user behavior, that is, the manner in which they produce and consume information?
2. How can we measure the informational, as well as the structural, characteristics of different social media?
3. Which social media are more appropriate for researchers to use, given their particular goals or tasks?

RESEARCH METHODS

Research Topics

We picked two topics pseudo-randomly to search for on multiple social media, with an exploration of people's views and opinions as the main research goal: net neutrality and the Thai coup d'état of May 2014. We also picked two semi-random topics to search with prediction as the main research goal: the predicted box office returns for the latest Star Wars movie, and the predicted stock price trends for Tesla Motors and Microsoft (both publicly traded companies). The details are described in Table 1.

The first topic, *net neutrality*, is a highly visible and current news topic. It refers to the principle of giving users "the right to use non-harmful network attachments or applications" and innovators "the corresponding freedom to supply them" (Wu, 2003). This topic is a controversial issue both in technical (Crowcroft, 2007) and political aspects (Cheng et al., 2010). The second topic refers to *the Thailand coup of 2014*, which was a minor international and political crisis that attracted some attention. Opinions and sentiments about those two topics can be observed and explored by people's expressions on social media.

Previous research suggested a predictive model of film box-office outcomes from Twitter data (Asur and Huberman, 2010). We also propose using data from multiple social media sources to predict the box-office outcomes of the 2015 movie, *Star Wars, Episode 7* (data was collected one year prior to the release of the film). Likewise, just as other research has attempted to predict stock market prices from data collected from social media (Gilbert and Karahalios, 2010), we suggest gathering data from multiple social media sources as a way to predict the price trends of two well-known stocks of public companies, *Tesla Motors* and *Microsoft*. Note that the purpose of this paper is not to execute these stated research tasks, but to show what a researcher might come across if he or she ventured to collect data from the four social media that we have selected for this study.

Social Media Selection

We chose to engage with four specific social media for our study. Twitter is a dominant social media service with more than 310 million monthly active users (as of Mar. 2016)[1]. YouTube is a video-sharing website with over 1 billion views per day[2]. The New York Times is a major newspaper with a very strong online presence. It receives over 78.1 million unique visitors per month[3]. Their website provides a feature that enables opinion expression and discussion through comments on various news articles. Reddit is a social networking site and news website where registered members can submit content and give "upvotes" and/or "downvotes" to others' postings. It has about 8.1 billion pages along with more than 29.5 million votes, as of Apr. 2016[4]. Reddit had over 243 million unique visitors in April 2016 . All of these social media have mechanisms to promote messages, be it through voting (Reddit, YouTube, NYT comments) or spreading messages by repeating them (Twitter).

Data Collection

[1] https://about.twitter.com/company

[2] http://www.youtube.com/yt/press/statistics.html

[3] http://nytmediakit.com/online

[4] http://www.reddit.com/about

Topic	Research Goal	Description
Net neutrality	Exploration	To find out range of views toward a particular controversial topic.
Thailand coup	Exploration	To find out opinions and sentiments of people to a political and social issue.
Star Wars	Prediction	To predict box office returns for the first weekend, month, and year, of "Episode 7,", per 2014 data (film was released in 2015).
Tesla Motors	Prediction	To predict the stock price trend over the next 2 months based on a variety of people's sentiments about the companies.
Microsoft	Prediction	To predict the stock price trend over the next 2 months based on a variety of people's sentiments about the companies.

Table 1: Description of Suggested Research Topics in the Study.

We interacted with application programming interfaces (APIs) for each social media we studied. This includes APIs for Twitter[5], YouTube[6], the NYT website[7], and Reddit[8]. The query terms we used are presented in Table 2. For Twitter searches, we chose relevant hashtags for specific topics. A hashtag is a word prefixed with the "#" symbol and usually represents a main theme of content. For YouTube searches, we simply used the query terms as shown. For the NYT searches, we used the query terms to retrieve articles, and then collected all user comments from these articles. For the Reddit data, we targeted specific subreddits that were appropriate for the topic, then we executed a search query on the topic as shown, and then retrieved the 10 most popular articles returned, and extracted the user comments.

The values shown in Table 3 represent the amount of data objects retrieved: tweets in the case of Twitter, video descriptions in the case of YouTube, and reader comments in the case of the NYT website and Reddit. We periodically and randomly sampled Twitter data over 7 days in May 2014. We also periodically sampled YouTube data in the same time period using a third party tool, which limited the amount of collected video descriptions to the hundreds of returns (YouTube only returns up to 1,000 results for every query). The Reddit and NYT comment data were each collected via API query once (per topic search).

RESEARCH FRAMEWORK

Richness of Information

In order to assess the "goodness" of the information retrieved, we looked to the literature on several specific measures regarding information acquired online and settled on three of them: the length of the content, the effect of having external links, and the effect of how frequently in time information is disseminated. We will call this collection of measurements the "richness of information."

Length of Content

The length of the content of is strongly correlated to the "readability" of an online review of goods and services (Ghose and Ipeirotis, 2011). We count the number of words in each data instance to measure this feature.

Presence of external links

Using the media richness theory, Jackson and Purcell (1997) proposed that hypertext allows communication products, such as Web pages, to indicate different levels of richness of information. The fact that a document has a unique resource locator (URL) link to external sites indicates that there are other relevant information sources to look into. This also increases the trustworthiness of the document because it provides a secondary source to the information within it.

Frequency of Information Dissemination

In studies done on Twitter activity in and around an earthquake occurrences in Japan, Sakaki et al. (2011) found that tweets increased in frequency when these events happened. The frequency of user engagement in social media strongly explains the helpfulness of online reviews (Ngo-Ye and Sinha, 2014). We see that the frequency of how often relevant information on a particular topic is disseminated on social media is an indicator of the popularity and the timeliness of the topic. It is also an indicator of how the particular social media is used.

Unique Relevant Word Use

Shah and González-Ibáñez (2011) defined unique coverage, as the unique Web pages visited during a collaborative effort, finding that synergic effect is dependent on users attaining unique coverage. Borrowing from this concept, we define the concept of unique relevant word use as the relevant words that came up in our search results that were exclusive to one social media or another. We argue that if a particular social media has a high amount of unique relevant word use, then it can provide the researcher with more diverse information on the topic being researched. This can also be an indicator of the kind of information found in particular social media data sources.

RESULTS AND DISCUSSION

Length of Content

[5] https://dev.twitter.com
[6] https://developers.google.com/youtube
[7] http://developer.nytimes.com/docs
[8] http://www.reddit.com/dev/api

Topic	Twitter	YouTube	NYT	Reddit
Net neutrality	*#NetNeutrality*	*net neutrality*	*net neutrality*	*net neutrality* in /r/politics
Thailand Coup	*#ThaiCoup*	*Thai coup* or *Thailand coup*	*Thailand coup*	*Thailand coup* in /r/worldnews
Star Wars	*#Starwars*	*Starwars episode 7*	*Starwars episode 7*	*Starwars episode 7* in /r/starwars
Tesla	*#Tesla*	*Tesla Motors*	*Tesla Motors*	Comments in /r/teslamotors
Microsoft	*#Microsoft*, *#MSFT*	*Microsoft*	*Microsoft*	Comments in /r/microsoft

Table 2: Query Terms used for Collecting Data.

Topic	Twitter	YT	NYT	Reddit	Total
Net neutrality	1,654	108	521	1,085	3,368
Thailand Coup	800	283	78	67	1,228
Star Wars	3,337	111	95	633	4,176
Tesla	1,700	104	274	124	2,202
Microsoft	3,196	224	618	87	4,125
Total	**10,687**	**830**	**1,586**	**1,996**	**15,099**

Table 3: Data Collected from Social Media.

We measured the number of words in the content of the data objects. The summary of these measures is given in Table 4 presents the . It is obvious that tweets show a relatively small number of words in them (less than 15, on average), given their 140-character limit. Reddit comments tended to be terse and succinct, averaging about 40 words per comment. By contrast, the descriptions of YouTube videos tended to have longer and more detailed explanations about the video. On average, these descriptions were about 122 words long. Likewise, The NYT article comments were about 103 words long, on average.

Presence of External Links

While the length of the content might indicate the amount of information a document has, it is not very appropriate in terms of extensiveness and credibility of information. Having an external link in a document gives a higher probability of that document's extensiveness and credibility. Figure 1 shows the probability of having external links in the documents across our selection of social media.

YouTube and Twitter content both show very high chances of having external links, on average those are 63% (SD = 14%) and 82.5% (SD = 5.8%), respectively. Both the NYT article comments and the Reddit comments showed much lower probabilities, which were, on average, 4.1% (SD = 1.6%) and 7.3% (SD = 3.9%), respectively.

Topic	Twitter	YouTube	NYT	Reddit
Net neutrality	15.0 (3.89)	132 (111)	87.1 (65.4)	51.9 (81.3)
Thailand Coup	16.4 (4.05)	160 (170)	220 (905)	41.5 (41.9)
Star Wars	14.7 (4.84)	119 (159)	62.7 (50.6)	28.8 (36.7)
Tesla	13.3 (4.82)	102 (135)	75.9 (65.0)	35.1 (32.6)
Microsoft	14.9 (4.17)	95.6 (126)	71.6 (61.1)	40.8 (44.4)
Average per SM	14.9 (1.10)	121.7 (25.7)	103.5 (65.8)	39.6 (8.57)

Table 4: Average Number of Words with S.D. in parentheses.

Frequency of Information Dissemination

The measure of how frequently data is posted on certain social media is an indicator of activity popularity of the topic. Table 5 illustrates the large distribution in this frequency that exists from one social media to another. For instance, while we can measure the frequencies of postings on Twitter in the low number of minutes (M = 1.33 minutes, SD = 0.73 minutes), the postings on YouTube appear around twice a month (M = 16.7 days, SD = 8.2 days) although this number drops to less than every 43 hours (SD = 36 hours), on average, when only looking at more recent postings (after 1/1/2014). On average, new NYT comments came up about once a day (SD = 30.7 hours), and new Reddit comments about once every 3.5 hours (SD = 3.0 hours).

Unique Relevant Word Use

Compared to Twitter and YouTube, the comments from the NYT articles and from Reddit show a very small number of frequently used words. While people use a limited number of words in tweets and YouTube descriptions, they have more flexibility when writing comments on news articles and con-

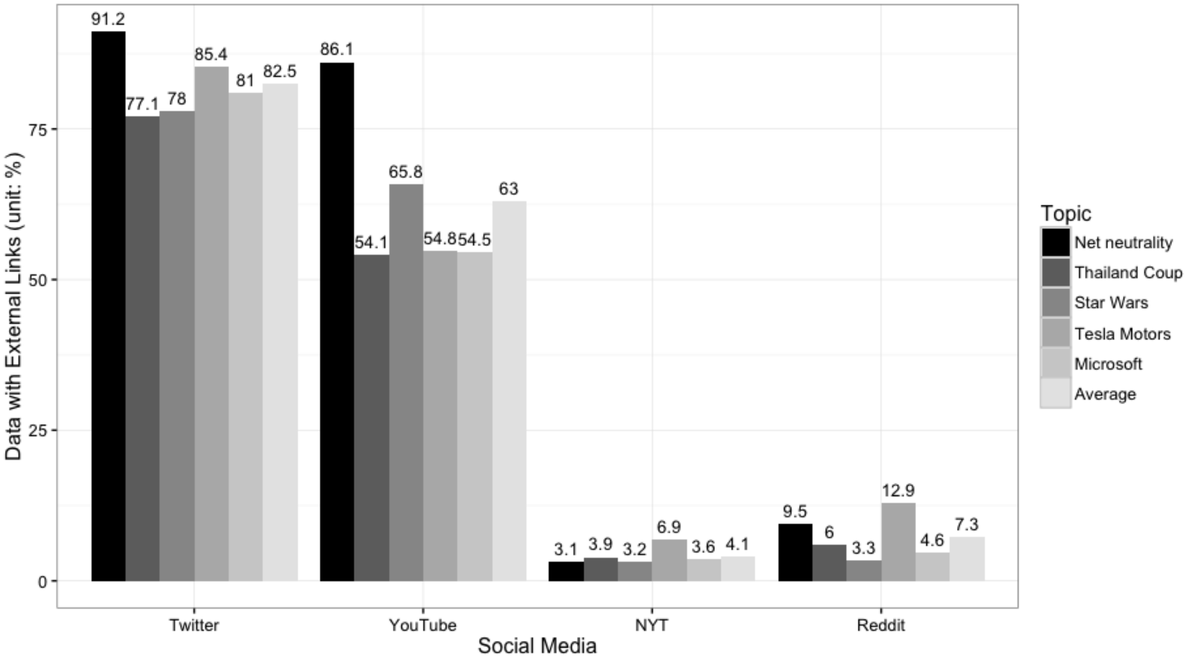

Figure 1: Percentage of Documents That Have External Links.

Topic	Twitter (secs)	YouTube (hours)	NYT (hours)	Reddit (mins)
Net neutrality	92.4 (216)	664 (3,147)	9.10 (89.1)	32.6 (44.8)
Thailand Coup	32.9 (173.7)	239 (2,994)	1.75 (535.2)	496 (125.5)
Star Wars	34.9 (155)	271 (1,543)	12.9 (40.9)	159 (264)
Tesla	130 (247)	560 (1,145)	77.8 (500)	95.6 (54.9)
Microsoft	110 (340)	273 (997)	33.4 (215)	278 (242)
Average	79.9s (44.0s)	401h (196h)	27.0h (30.7h)	212m (183m)

Table 5: Frequency of Information Generation.

troversial arguments that they are interested in, as in the case of comments found on the Reddit and NYT articles. For instance, words like “people,” “like,” “will,” and “just” appear more frequently in the comments of the NYT and Reddit articles, than in tweets and YouTube descriptions.

Also, words like “new” and “follow” appear almost exclusively on a list of frequent YouTube words. References to other social media, such as Twitter and Facebook, appear exclusively amongst the YouTube frequent word lists. This suggests that, of all the social media we studied, YouTube served as a common bridge between users of different social media.

The words “new” and “love” are popular with Twitter users. This agrees with a common view that Twitter messages reflect user behavior that clusters in two main areas: expressing personal opinions and information sharing (Naaman et al., 2010).

Limitations

The way we collected the data related the suggested topics might have introduced biases. For instance, while tweets were collected through particular keywords (*hashtags*), data from YouTube was collected through keyword searches. In Reddit, we focused on some sub-reddits looking for relevant content with keywords of interest. We used the data collection strategies considering the different use cases of each social media, but it may bring us different results if we strictly use the same approach to all social media and topics. To address this problem, we plan to apply a variety of different ways of gathering data of interest in future studies to see to what extent data collection methods influence the results in a comparative analysis. The different volumes of data that we retrieved from each platform may be difficult to compare against one other because of innate platform or media differences.

CONCLUSION & FUTURE WORK

We presented a result of a comparative study of multiple social media - Twitter, YouTube, the New York Times website comments, and Reddit comments - collecting and analyzing data gathered for our suggested research goals and topics. Our analysis shows the trade-offs between different social media services and provides a suggested framework to compare cross-platform data collection and analysis utility.

The criteria we used in this study may not be comprehensive. In fact, the increasing number of social media along with their different data formats makes research like this more difficult. In our future work, we would like to examine other aspects of social media that we did not address here, for example, how easily a particular social media allows its users in general, and researchers in particular, access to its data, and how social networking features on social media might influence the type of data that researchers collect.

ACKNOWLEDGMENTS

The work reported here is supported by the National Science Foundation (NSF) BCC-SBE award no. 1244704 and the Institute of Museum and Library Services (IMLS) grant no. RE-04-12-0105-12.

References

Asur, S. and Huberman, B. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, pages 492–499.

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an Influencer : Quantifying Influence on Twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62. ACM.

Cheng, H. K., Bandyopadhyay, S., and Guo, H. (2010). The Debate on Net Neutrality: A Policy Perspective.

Crowcroft, J. (2007). Net neutrality: the technical side of debate: a white paper. *ACM SIGCOMM Computer Communication Review*, 37(1):49.

Ghose, A. and Ipeirotis, P. (2011). Estimating the helpfulness and economic impact of product reviews : mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Gilbert, E. and Karahalios, K. (2010). Widespread Worry and the Stock Market. *ICWSM*, pages 58–65.

Hawn, C. (2009). Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health affairs (Project Hope)*, 28(2):361–8.

Hughes, D. J., Rowe, M., Batey, M., and Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569.

Jackson, M. and Purcell, D. (1997). Politics and media richness in world wide web representations of the former Yugoslavia. *Geographical Review*, 87(April):219–239.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Liu, Y. (2006). Word of Mouth for Movies : Its Dynamics and Impact on Box Office Revenue. *Journal of marketing*, 70(3):74–89.

Loader, B. D. and Mercea, D. (2011). Networking democracy? *Information, Communication & Society*, 14(6):757–769.

Naaman, M., Boase, J., and Lai, C.-h. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, pages 189–192, New York, New York, USA. ACM.

Ngo-Ye, T. L. and Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, 61:47–58.

Sakaki, T., Toriumi, F., and Matsuo, Y. (2011). Tweet trend analysis in an emergency situation. *Proceedings of the Special Workshop on Internet and Disasters - SWID '11*, pages 1–8.

Shah, C. and González-Ibáñez, R. (2011). Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 913–922. ACM.

Shirky, C. (2011). Political Power of Social Media-Technology, the Public Sphere Sphere, and Political Change. *Foreign Aff.*, 1:28–41.

Wallsten, K. (2010). "Yes We Can": How Online Viewership, Blog Discussion, Campaign Statements, and Mainstream Media Coverage Produced a Viral Video Phenomenon. *Journal of Information Technology & Politics*, 7(2-3):163–181.

Wu, T. (2003). Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology Law*, 2:141.

Yang, Z., Jones, I., Hu, X., and Liu, H. (2015). Finding the right social media site for questions. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 639–644. ACM.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing Twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, Berlin Heidelberg.