Information Sources Cited and Relayed in Political Conversations on Twitter

Jung Sun Oh[1], Jae-wook Ahn[2], Jisue Lee [3]
[1]University of Pittsburgh
[2]IBM TJ Watson Research Center
[3]Florida State University

Abstract
Using the Twitter data collected prior to the Presidential Election in Korea in 2012, we ask questions regarding influential sources of information in public political discourse. The frequently cited sources, being included as URLs in political tweet messages, are identified and categorized. The result shows that people rely on various sources of information beyond the traditional news media, but the pattern of sharing differ by sources.

Keywords: Twitter; political communication; information sources; retweet
doi: 10.9776/16591

Contact: jsoh@pitt.edu

1 Introduction

Over the past years, Twitter has gained attention as a platform for political discourse. In this study, we aim to uncover the kinds of information sources people bring into their political conversation and the relative importance of different sources, by looking at the URLs included in political tweets. More specifically, using the Twitter data collected during the campaign period of the Presidential Election in Korea in 2012, we explore the following research questions in this paper.

- RQ1: What kinds of information sources are invoked in political discourse?
- RQ2: To what extent do different sources of information gain attention and get endorsed (in the form of retweeting) by ordinary citizens?

Citing a URL in Twitter message is an attempt to draw attention to and/or raise awareness of the information therein (be that news, facts, ideas, opinions, and so on). We posit that in aggregate it reflects people's assessment of the relevance or importance of information sources to the matter at hand.

Traditionally, mainstream media has been believed to bear a dominant influence on the shaping of political discourse, especially in the electoral period. The capability of spreading campaign propaganda to a large audience, often selectively, creates such influence. One of the premises of social media, especially Twitter (with its openness), is to liberate the flow of information and to harness opinions of crowds. The extent to which various media outlets and information sources are cited in Twitter messages regarding the candidates and the campaigns can be a gauge determine whether the premise holds in actual election settings.

2 Background & Related Works

Increasingly, more and more citizens rapidly adopt the Internet and its related technologies to engage in social communication and interaction with others for political purposes. Social media platforms such as Twitter, Facebook, and Youtube are considered emerging virtual public spheres where individuals freely discuss public affairs with others and form public opinion (Papacharissi, 2002; Shirky, 2011). Since 2008 more than half of the voting population in the U.S. have proactively participated in political discourse by posting political thoughts/comments and sharing political news articles with others during election campaigns (Smith, 2009; Rainie et al., 2012).

Given the nature of rapid information distribution and dissemination, Twitter is considered as a powerful mass communication as well as information diffusion tool (Suh et al., 2010, Jansen et al., 2009). Particularly, the retweeting feature well serves information diffusion among the networked individuals on Twitter. Expanding information access to a variety of information sources is critical to enhancing open political discourse and deliberation (Papacharissi, 2002; Shirky, 2011).

3 Data & Analysis

3.1 Data Collection

The Twitter data used in this study was collected prior to the Presidential Election in Korea in 2012. From April 2 to December 21, 2012 (two days after the Election Day), we collected tweets using the names of three presidential candidates as keywords: Geun-hye Park, Jae-in Moon, and Cheol-soo Ahn. The Python Twitter API named Twython (https://github.com/ryanmcgrath/twython) was used to access Twitter REST Search API (https://dev.twitter.com/docs/api/1.1/get/search/tweets). In total, approximately 13.5 million tweets were downloaded, including 1,661,422 unique tweets and the retweets of those unique tweets.

3.2 URL Processing

The first step of the analysis was to extract URLs in the collected tweets. We identified URLs in tweets using regular expressions. Out of 1,661,422 unique tweets in our dataset, 933,169 (56%) include URLs. URLs included in Twitter message are typically shortened (using services like t.co or bit.ly). The shortened URLs were first resolved to the full URLs. In many cases, a URL in tweets turned out to be a shortened version of another previously shortened URL. In those cases, we attempted to restore the original URL by recursively resolving them. Of the 929,491 shortened URLs found in our dataset, 868,782 were successfully restored to their original (unshortened) form. In the end, 239,890 unique URLs were extracted.

3.3 Source Categories and Coding

In order to categorize information sources represented as URLs, we constructed a codebook for the domains of the URLs appeared in tweets. We started with the initial set of codes representing different types of news media. The initial list of domains belonging to those media categories was obtained from Nielsen Korea. The rest of the codebook was built through an iterative process. We sorted the URLs by domains and by frequencies of their occurrence in our dataset. For unknown/uncoded domains, we manually visited the URLs to determine the code appropriate for the nature and content of the domain in question. If needed, a new code was added to the codebook. Those domains that are no longer available and those that appeared only a few times in tweets were left out.

In the end, the codebook contained 1024 domains categorized into 16 different codes. Among those, excluding two codes assigned to third party Twitter application sites that are not relevant to the purpose of this study, 14 codes are used in the analysis. Table 1 shows the category of sources and the corresponding codes. Having constructed the codebook, we assigned a category code to each URL based on its domain.

Macro category	Micro category	Description
News media	IN-TM	Internet news — by mainstream news publishers (Internet version of traditional newspapers)
	IN-AM	Internet news — by alternative news publishers (published online only)
	TN-TM	Televised (broadcast) news — by mainstream TV stations
	TN-AM	Televised (broadcast) news — by Internet stations (streaming)
	WZ	News magazine
	PN	Portal news
	FN	Foreign news/magazine
Community forums	CF	Various user communities, discussion boards, forums
Social media	SM	Video sharing sites (e.g. YouTube), image sharing sites (e.g. Flickr, Instagram), social networking sites (e.g. Facebook), …
Podcast	PC	Podcast
Blog	BG	Personal / group blog
Websites	WS-P	Official political websites (e.g. political party, campaign website)
	WS-X	Partisan websites
	WS-O	Other websites

Table 1. Source categories and descriptions

3.4 Metrics of Spread

We define three major metrics for this paper: (1) URL dissemination rate by retweeting, (2) URL lifespan, and (3) URL survival rate. Dissemination rate is defined as the number of retweets per each URL. It identifies how frequently a specific URL is retweeted after it is introduced in a tweet. URL lifespan is the duration of time while a URL is active within retweet chains. Here it is calculated by “most recent RT timestamp” – “oldest RT timestamp” (days). Survival rates are defined as the fraction of retweeted URL counts within a specific time period out of total counts.

4 Results

4.1 RQ1 – what kinds of sources?

Table 1 and 2 summarize tweet statistics by categories. News media (IN or TN) is the most frequently appeared category, with a large number of unique URLs and a high dissemination rate. In the realm of Twitter, Internet News (IN) clearly outweighs Televised News (TN), showing 15 times more URLs and 22 times higher RT counts. Moreover, among Internet News, those categorized as alternative media news (IN-AM) were introduced and retweeted more frequently than the traditional media news (IN-TM). In case of Televised News (TN), however, alternative media (TN-AM) shows a much smaller number of URLs and retweet counts than mainstream media.

Micro category	# of URLs	RT count	Dissemination rate	RT standard deviation
BG	3,413	112,061	32.8	162.3
CF	11,875	254,050	21.4	94.2
FN	11	887	80.6	95.0
IN-AM	36,450	1,209,685	33.2	177.2
IN-TM	21,221	756,289	35.6	203.5
MZ	284	5,564	19.6	48.5
PC	49	1,485	30.3	107.6
PN	34,449	937,561	27.2	139.3
SM	8,637	625,268	72.4	305.4
TN-AM	969	23,126	23.9	162.8
TN-TM	2,862	64,762	22.6	115.9
WS-O	750	27,701	36.9	182.6
WS-P	3,749	122,622	32.7	143.8
WS-X	2,557	93,643	36.6	151.0

Table 2. Micro category tweet statistics

Macro category	# of URLs	RT count	Dissemination rate	RT standard deviation
Websites	7,056	243,966	34.6	151.0
Social media	8,637	625,268	72.4	305.4
Community forums	11,875	254,050	21.4	94.2
News media	95,952	2,991,423	31.2	169.5
Blog	3,413	112,061	32.8	162.3
Podcast	49	1,485	30.3	107.6

Table 3. Macro category tweet statistics

4.2 RQ2 – the extent of spreading (RT) by category

Macro category	Max	Min	Average	Standard deviation
Websites	262	0	2.9	13.3
Social media	262	0	6.4	25.7
Community forums	263	0	1.3	8.2
News media	263	0	1.7	11.2
Blog	245	0	4.4	21.1
Podcast	231	0	20.0	49.8

Table 4. URL lifespan (Days)

In addition to the average dissemination rate by category, shown in Table 2 and Table 3, the average life span and the survival rate of URLs in different categories are calculated. 86% of the unique URLs in our dataset have lifespans of less than one day (205,432 out of 239,890). Table 4 compares URL lifespans by category. Podcasts has the longest lifespan (20 days) in Table 4 whereas the number of URLs and retweets are relatively much smaller (Table 2 and 3). Social media sites and blogs survived, on average, 6.4 and 4.4 days respectively. News media sources' lifespan is less than two days on average. It reflects the nature of news media where recency is more important than other media.

Figure 1 shows the distribution of URL lifespan. It was assumed that it would follow a typical power-law distribution, considering the fact that most of the unique URLs (86%) survived less than one day. Figure 1 coincides with the assumption until around the duration (lifespan) of 100 days, but after the day 100 there is no more steep decrease in numbers and even a small spike at around the day 250. Therefore we broke down the distribution by categories and calculated survival rates in three time periods: less than one day, equal or greater than 100 days, and equal or greater than 200 days (Table 5). It revealed clear differences in survival rate by categories. For instance, while more than 12% of podcasts stayed circulating past 100 days (>=100 and >=200), only 0.16% of new media survived longer than 100 days. Blogs and social media also exhibit higher survival rates in the longer terms but not as high as the podcasts.

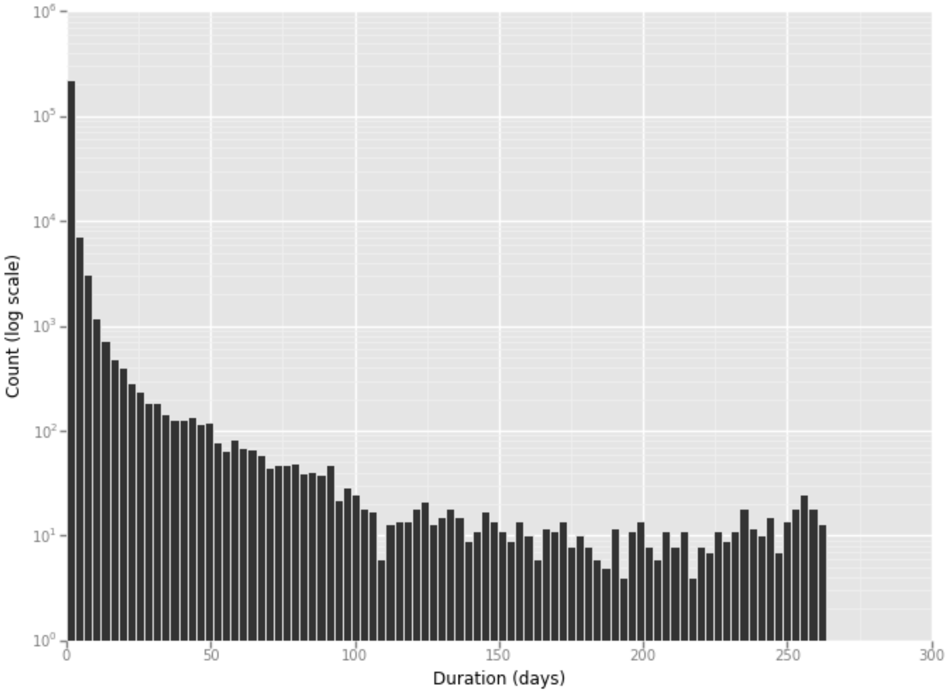

Figure 1. Lifespan distribution

Macro category	Lifespan < 1	Lifespan >=100	Lifespan >=200
Websites	70.46%	0.34%	0.13%
Social media	70.29%	1.57%	0.64%
Community forums	83.78%	0.13%	0.03%
News media	82.06%	0.28%	0.11%
Blog	76.88%	1.08%	0.44%
Podcast	67.35%	10.20%	2.04%

Table 5. Tweet survival rates by categories in three time periods

5 Discussion & Conclusion

The purpose of this study was to understand how people cite and relay information from different sources in Twitter, especially in the context of political communication. Using a large scale Twitter data, we identified the main categories of information sources cited in Twitter by looking at the URLs and also examined the extent of spread and longevity of those sources. The result shows that people rely on various sources of information beyond the traditional news media, but the pattern of sharing differ by sources.

As a preliminary study, the analysis of citing and relaying patterns in this study is admittedly simple and descriptive, yet the results show a few points worth noting. First, comparing news media, it is notable that alternative media outlets with an Internet-only presence are cited far more often than traditional mainstream news publishers. While the relatively large number of such alternative news sites may have contributed to the observed difference, it still testifies the sizable role and potential influence of alternative media in political discourse. Second, people actively share the contents from sources other than news media, including social media, community forums, and blogs. Although smaller in numbers, these sources tend to stay longer on Twitter. For instance, the average lifespan of social media contents is more than three times longer that that of news media.

In this study, in looking at the rates of spread and survival of cited information (URLs), we only considered the categories of source. Undoubtedly, the actual content of cited articles would be an important factor, and the question of who participated in spreading the information would also be relevant considering the networked nature of Twitter. We plan to incorporate these factors in our future study.

6 References

Jansen, B. J., Zhang, M., Sobel, K., & 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.

Papacharissi, Z. (2002). The virtual sphere: The Internet as a public sphere. *New Media Society, 4*(1), 9-27. doi: 10.1177/14614440222226244

Rainie, L, Smith, A., Scholozman, K. L., Brady, H., & Verba, S. (2012). *Social media and political engagement*. Retrieved from Pew Research Center: http://pewinternet.org/~/media//Files/Reports/2012/PIP_SocialMediaAndPoliticalEngagement_PDF.pdf

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign Affairs, 28(January/February).* Retrieved from http://www.bendevane.com/FRDC2011/wp-content/uploads/2011/08/The-Political-Power-of-Social-Media-Clay-Sirky.pdf

Smith, A. (2009). *The Internet's Role in Campaign 2008*. Retrieved from Pew Research Center: http://pewinternet.org/~/media//Files/Reports/2009/The_Internets_Role_in_Campaign_2008.pdf

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweets in Twitter network. *Proceedings of IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust*. doi:10.1109/SocialCom.2010.33