Asian Journal for Public Opinion Research - ISSN 2288-6168 (Online)
Vol. 5 No.4 August 2018:-285-301
http://dx.doi.org/10.15206/ajpor.2018.5.4.285

Analyzing Public Opinion with Social Media Data during Election Periods: A Selective Literature Review

Jin-ah Kwak[1]

KAIST, South Korea

Sung Kyum Cho

Chungnam National University, South Korea

Abstract

There have been many studies that applied a data-driven analysis method to social media data, and some have even argued that this method can replace traditional polls. However, some other studies show contradictory results. There seems to be no consensus as to the methodology of data collection and analysis. But as social media-based election research continues and the data collection and analysis methodology keep developing, we need to review the key points of the controversy and to identify ways to go forward. Although some previous studies have reviewed the strengths and weaknesses of the social media-based election studies, they focused on predictive performance and did not adequately address other studies that utilized social media to address other issues related with public opinion during elections, such as public agenda or information diffusion. This paper tries to find out what information we can get by utilizing social media data and what limitations social media data has. Also, we review the various attempts to overcome these limitations. Finally, we suggest how we can best utilize social media data in understanding public opinion during elections.

Keywords: social media analysis, election prediction, public opinion

[1] All correspondence concerning this article should be addressed to Jin-ah Kwak at KAIST, 291 Daehak-ro, Eoeun-dong, Yuseong-gu, South Korea or by email at jinah.kwak@kaist.ac.kr

Introduction

In the past 10 years, there have been ample arguments suggesting that data-driven analysis methods using social media data can supplement or even replace traditional polls and surveys. The affordability and timeliness of data collected from social media is appealing. Some researchers are interested in investigating the political discourse spread in social media and argued that social media data has potential utility as an indicator of political opinion and are comparable to offline surveys (Caldarelli et al., 2014; Chung & Mustafaraj, 2011; DiGrazia, McKelvey, Bollen, & Rojas, 2013; Kalampokis Tambouris, & Tarabanis, 2013; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; O'Leary, 2015). However, other scholars have suggested that the majority of the social media messages are "pointless babble" and that there is little consensus regarding methodology and evaluation (Boyd & Crawford, 2012; Couper, 2013; Gayo-Avello, 2013; Jungherr, Jürgens, & Schoen, 2012; Metaxas, Mustafaraj, & Gayo-Avello, 2011; Murphy et al., 2014; Ruths & Pfeffer, 2014; Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016; Schoen et al., 2013; Tufekci, 2014; Weller, 2015; Yu, 2012). There is no consensus on the validity of this method.

Many researchers tried to overcome the limitations by improving sampling or weighting, or by reducing data noise (Baldwin, Cook, Lui, MacKinlay, & Wang, 2013; Barberá, 2016; Chang, Rosenn, Backstrom, & Marlow, 2010; Choy, Cheong, Laik, & Shung, 2011; Choy, Cheong, Laik, & Shung, 2012; Davis, 2017; Diaz, Gamon, Hofman, Kıcıman, & Rothschild, 2016; Flemming & Sonner, 1999; Kalampokis et al., 2013; Karimi, Wagner, Lemmerich, Jadidi, & Strohmaier, 2016; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011; Tufekci, 2014;Wang, Rothschild, Goel, & Gelman, 2015; Zhang, Ram, Burkart, & Pengetnze, 2016).

In this paper, we will review methodologies that have been widely used to predict election results and to evaluate public opinion. Next, we will point out the limitations of those methodologies and then introduce the efforts to overcome those limitations. Finally, we would like to propose a way to measure the direction of public opinion through big data analysis. Specifically, we will focus on how to use social media analysis methods and argue that we can utilize them in understanding agenda, rather than in election forecasting. We pay attention to the possibility of using social media data for understanding public opinion, rather than predicting election outcomes.

Review: Previous Work

Previous review papers only focused on the accuracy of predictions. For example, some researchers (Gayo-Avello, 2013; Schoen et al., 2013; Gayo-Avello, 2011, 2012a, 2012b; Goldstein & Rainey, 2010) review and make general arguments about how methodology led to the success or failure of social media predictions. Also, another researcher (Phillips, Dowling, Shaffer, Hodas, & Volkova, 2017) reviewed the latest literature in 2017 and focused on larger scope predictions such as stocks and marketing or public health, etc. Those review papers only focused on the success or failure of results.

For this paper, we first conducted a database search in June 2018 on Google Scholar for articles published since 2010, which include the terms "social media" and "election prediction." Then, we tracked each paper to include papers that attempt to use social media data to predict offline political phenomena. Our search criteria yielded 534 articles, of which we selected 69 that matched our research topic.

Methods Used in Predictive Modeling

There are three major predictive modeling methods for election prediction using social media data. Those three methods are used in various research papers and by big data analysis companies. .

Counting volume

The oldest and a basic attempt to analyze public opinion using social media data is counting volume, which simply counts the number of a specific word's appearances. At the beginning of social media research, most studies used Twitter, which was the most popular at that time. This method has been widely used since the correlation between the volume of social media and the election results was discovered. Counting the number of tweets which contain a reference to a political party or a politician has been used in many studies (Khatua, Khatua, Ghosh, & Chaki, 2015; Skoric, Poor, Achananuparp, Lim, & Jiang, 2012, Tumasjan, Sprenger, Sandner, & Welpe, 2010; Williams & Gulati, 2008). After counting tweets or expressions, the ratio of each political party or candidate on social media is compared to the actual election results, or the ranking by social media volume and the ranking by survey or by share of voters are compared.

These researchers insist that the relative volume of tweets mirrors the results of the election closely. To make a clear difference, most of the researchers calculate the Mean Absolute Error (MAE), which has been used to compare the accuracy of political

information on social media relative to election polls. However, there is no theoretical explanation about the relationship between the volume of tweets and election choice.

Sentiment analysis and machine learning

The basic concept is similar to counting, but this method adds sentiment of word or location information as a variable. Then it uses regression or machine learning analysis (Ceron, 2014) instead of simple counting. Sentiment analysis basically measures the affirmation and the negation of extracted words by morpheme analysis. Analysis of sentiments or opinions is assumed to be better than volume counting. It extracts and analyzes opinion-oriented text, recognizing positive and negative opinions, and quantifying how positive and negative entities are (Chen, Wang, & Sheth, 2012; Pimenta, Obradovic, & Dengel 2013; Sang & Bos, 2012). To define which word has what sentiment, usually a lexicon-based sentiment analysis dictionary, or LIWC (Linguistic Inquiry and Word Count) tool is used. More recently, machine-learning methods are used to detect the sentiment of text.

However, sentiment analysis is still not perfect. Though it seems more sophisticated than the method of measuring volume, this method has weaknesses. The sentiment of the words depends upon context. So, the same words may have different sentiments but the computer is not yet perfectly able to find this sentiment. Some researchers also pointed out that ironic and sarcastic expressions are very frequent and difficult to detect (Reyes & Possp, 2014; Clavel & Callejas, 2016)

Trends: Search keywords

In a more recent approach, researchers use Google Trends, a service that shows how often a specific word is searched by region and time. Unlike social media posts, search keywords are not shared with others. So, they can be said to reflect the user's honest thoughts. Google Trends analysis is a relatively easy tool to use, and so, not only researchers but also individuals or reporters can easily approach and analyze this data. Moreover, not only words like a political party or the name of a politician, but also more political issues related to specific politicians can be analyzed (Stephens-Davidowitz, 2014).

However, it is difficult to measure opinions toward candidates on the basis of search frequencies of candidates. For example, we can think of what you searched for as an interest, but because of instances of searching for criticism or out of curiosity, it does not link to support or willingness to vote for a candidate.

Limitations of Predictive Modeling

Many researchers point out that monitoring what users share or search for on social media and on the web has led to greater insights into what people care about or pay attention to at any moment in time. However, social media and search results can be readily manipulated, which is something that has been underappreciated by the press and the general public. We will describe and summarize the limitations of using social media data.

Crawling and platform difference

There are various social media platforms such as Twitter, Facebook, Instagram, blogs, Google Plus, etc. We should focus on platform-specific biases. Every social media platforms has its own data collection and accessibility policies. Also, data provision policies for research are different. For example, Twitter provides an API (Application Programming Interface) to provide tweet data. However, according to each API and their required conditions, researchers might obtain a biased sample from Twitter. Twitter provides a glance into its millions of users and billions of tweets through a "Streaming API," which provides a sample of all tweets matching some parameters preset by the API user. However, the essential drawback of Twitter API is the lack of information concerning what and how much data users get. This leads researchers to question whether the sampled data is a valid representation of the overall activity on Twitter. As an example, some researchers (Ghosh et al., 2013; Morstatter, Pfeffer, & Liu, 2014; Morstatter, Pfeffer, Liu, & Carley, 2013; O'Connor et al., 2010) acquired different data collected during the same period due to their collection method.

Also, it is obvious that different social media platforms are used in different countries. While popular social media varies from country to country, there are a number of studies that do not take this into consideration. For example, at the time of the US presidential election in 2016, the mainstream media in the US did not predict Trump's victory. However, Google Trends expected Trump to win early because the number of Trump searches was greater than the number of Clinton searches. Could this result be applied to other countries? It is true that Google Trends successfully predicted (Lui, Metaxas, & Mustafaraj, 2011; Metaxas, Mustafaraj, & Gayo-Avello, 2011) the US presidential election. However, the situation in Korea and the US is different. According to Naver, Naver ranked first in the PC search market in Korea in March this year with 75.4% of searches, while Google had only 6.7%. Google Trends only reflects the search results from Google, making it less representative of the Korean search market (Traffic Difference: US vs. Korea, July 2018, Alexa.com).

Representativeness: Biased Sample

The preceding problems lead to biased samples. The main challenge is the securing of the representativeness of social media data. Just having a large number of tweets does not mean that there has been a representative sampling of the voting population. A common assumption underlying many large-scale social media-based studies of human behavior is that a large enough sample of users will drown out noise introduced by peculiarities of the platform's population. In modern opinion polling, insuring a representative sample is a core issue, that is, each individual in any particular target population should have a greater than zero probability of being sampled (Barberá & Rivero, 2015; Chen et al., 2012; Gayo-Avello , 2011; Malik, Lamba, Nakos, & Pfeffer, 2015; Ruths & Pfeffer, 2014; Tufekci, 2014).

However, biases may vary across different social media platforms. For instance, in the US, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents," whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (Ruths & Pfeffer, 2014). Such factors can affect the results of social media prediction analysis. Representativeness is a crucial problem of social media, and so many researchers in various fields, like social science and computer science, consistently point out the limitations.

Validity of methods: Irreplicability and incomparability

There is no agreement among researchers yet on the criteria for successful prediction. Metaxas et al. (2011) test the predictive power of social media in several senate races of two recent US Congressional elections. They review the findings of other researchers and try to duplicate their findings both in terms of data volume and sentiment analysis. They described three necessary standards that any theory aiming to predict elections competently and consistently using social media data should follow: 1) the prediction theory should be an algorithm with carefully predetermined parameters, 2) the data analysis should take into consideration the difference between social media data and natural phenomena data, and 3) it should contain some explanation about why it works.

In most cases, researchers have filtered their data on the basis of decisions clearly made after the elections were over and the results were known (including which parties' tweets were included). This has led to an inability to replicate reported success rates (Gayo-Avello, 2013; Marchetti-Bowick & Chambers, 2012).

Overcoming the Limitations

Recently, there have been various efforts to address the issues identified above. Here, we will present attempts to solve the most discussed problems of representativeness and other analytical methodologies that can utilize social media.

Sampling: User demographics

An important limitation in previous studies of political behavior using Twitter data is the biased sample. There are differences of demographic composition between web panels and the population of the U.S. in the 2014 election (Scott et al., 2015) and between search and Twitter data in the 2012 presidential election in the U.S (Diaz et al., 2016). Many researchers address this challenge by developing new machine learning methods that will allow researchers to estimate the age, gender, and race of any Twitter user in the U.S. with high accuracy. A variety of methodologies are used, including classifying and predicting users through machine learning, using all the variables available online as input data (Chang et al., 2010; Karimi et al., 2016; Mislove et al., 2011).

After inferring the demographics of social media users, they apply weighting techniques to predict the percentage of votes that individual candidates will receive (Barberá , 2016; Choy et al., 2011; Choy et al., 2012; Davis, 2017; Diaz et al., 2016; Flemming & Sonner, 1999; Gayo-Avello, 2011; Wang, 2015). We can match the demographic property distribution of the population, or apply weighting according to census data to correct the results. For example, some researchers (Diaz et al., 2016; Gayo-Avello, 2011) weighted the demographic distributions of social media users according to the census. On the other hand, random sampling (Barberá, 2016) or quota sampling (Davis, 2017) has also been used in sample construction.

Data and method

There have been many studies to develop methods of analysis and data collection. Some researchers (Kalampokis et al., 2013; Tufekci, 2014) insist that data filtering is an important step. For example, when crawling social media data, researchers collect social media posts containing keywords or hash tags that are relevant to the research topic. If the wrong keyword is selected at this time, meaningful data may be discarded or meaningless data may be collected, which generates data noise. Advanced natural language processing algorithms have been developed to reduce noise not only in election predictions, but also in other predictive studies; however, it is still a difficult task (Baldwin et al., 2013; Zhang et al., 2016).

Suggestions

There have been attempts to analyze the structure or diffusion process of public opinion. These attempts tried to show how opinions are diffused or how media messages are structured. We believe that these methods are more appropriate for understanding public opinion, and we think social media analysis can be a way to understand public opinion during elections, rather than predict election results.

Exploring Information Structure by Network Analysis

Tweets are propagated in various ways. Retweeting is the most effective, as it can potentially reach the most people, given its viral nature (Petrovic, Osborne, & Lavrenko, 2011). Retweeting is the action of reposting someone else's tweet inside your own message stream, and there are generally two ways to do it on Twitter. Users can either manually edit the original tweet or add "RT @userA" (or something similar) to indicate that the original tweet came from userA, or they can use a retweet button, which does not allow them to change the original tweet. In Twitter's API, the tweet-retweet connection is marked. In short, a RT network is a message propagation network. We can interpret who spread whose message. Also, by network analysis, we can detect the important issues spread in Twitter (Cameron, Barrett, & Stewardson, 2016; Dokoohaki, Zikou, Gillblad, & Matskin, 2015; Petrovic et al., 2011). By analyzing keyword networks, which are frequently used keywords in Twitter, we can calculate a clustering coefficient value that shows some information can be spread widely, and some information cannot be spread and remains isolated.

Finding Topics and Frames by News Article Analysis

Traditionally, researchers read a random sample of articles and then categorize or analyze the tone of the article, which is called content analysis by human coder. Mining public opinion from news articles is a traditional area of opinion analysis. Through news articles, we can measure the direction of public opinion. As computer analysis technology combines with content analysis, it has become possible to analyze news articles and reports on a large-scale (An & Gower, 2009; Krstajić, Mansmann, Stoffel, Atkinson, & Keim, 2010; Sjøvaag, H., & Stavelin, 2012). Automated analysis techniques, such as clustering and classification techniques, can be applied to news articles. Usually, by counting the word frequency for a document set then putting them in a vector, we can analyze and apply co-occurrence words analysis (Chen, 2009; Li et al., 2013), topic detection (Papadopoulos et al., 2014; Wang et al., 2008), and discourse framing analysis (Dimitrova et al., 2005;

Gamson, 1989; Semetko & Valkenburg, 2000; Tian & Stewart, 2005). By analyzing news articles, we can understand the discourse constituted in the media and measure the direction of public opinion.

Understanding Social Media User Behavior

Each social media platform is characterized by its ability to analyze various user behaviors. Especially on Twitter, three major activities can be interpreted differently. Other researchers (Stieglitz & Dang-Xuan, 2012) and the political blog FiveThirtyEight (Roeder, Mehta, & Wezerek, 2017) suggest a ternary plot to measure the social media presences of some of the most powerful politicians in the United States. Tweets toward the top have a higher share of retweets, those toward the bottom right have a higher share of likes, and those toward the bottom left are in the ratio danger zone - a higher share of replies. Facebook is a more personalized social media, allowing the analysis of friendships, private settings, likes, and favorite pages. Facebook allows an easy measurement of the performance of a Facebook page. Obtaining and using this data is becoming more and more difficult due to privacy concerns. We can track likes, page views, reach, and more, relatively easily. As mentioned above, there are many studies analyzing different user behavior by different social media platforms, for example, YouTube's likes, liked channels, or Instagram's likes, replies, tags, etc.

Discussion

There has been a conflict between positive and negative opinions about research methods and results that use social media. There is controversy about the scientific outcomes of research methods and the generalization of research results, but the claim of predicting the election results with social media data is not well supported. Despite criticism of the usefulness and predictability of social media data, social media research is constantly being attempted due to its advantages such as autonomy, immediacy, and size of data.

In this study, we examined the usefulness of social media analysis for analyzing public opinion formation, propagation process, and media issues. In order to solve the problem of methodological issues and representativeness, researchers try to overcome these challenges through sampling and data collection efforts, but the emergence of bot and spam accounts has heightened the controversy about the manipulability and reliability of social media data (Chu, Gianvecchio, Wang & Jajodia, 2010; Haustein et al., 2016;

Morstatter, Dani, Sampson & Liu, 2016; Ratkiewicz et al., 2011; Shu et al., 2017). The methodological errors that still need to be resolved have not been overcome sufficiently.

Therefore, we propose using social media data to understand agendas rather than to use them for forecasting. In other words, we suggest utilizing the methods of network analysis and news content analysis that we introduced. Rather than forecasting the election results, it is reasonable to identify which candidates' campaigns or issues are agenda-setting through the media, and to identify public interest in the agenda. Due to the problems such as data noise and sample bias, it is too risky to predict behaviors. Instead, it is possible to analyze the issues and the agenda of a specific politician or political party. Therefore, taking advantage of the benefits of social media data, identifying public agendas, and measuring interest will be more relevant and useful in grasping public opinion in the future.

References

An, S. K., & Gower, K. K. (2009). How do the news media frame crises? A content analysis of crisis news coverage. *Public Relations Review, 35*(2), 107-112. DOI: 10.1016/j.pubrev.2009.01.010

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text, how diffrnt social media sources?. *In Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 356-364).

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users*. Social Science Computer Review, 33(6)*, 712-729. DOI: 10.1177/0894439314558836

Barberá, P. (2016). Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. Working Paper.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon*. Information, communication & society, 15(5)*, 662-679. DOI: 10.1080/1369118X.2012.678878

Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., & Riotta, G. (2014). A multi-level geographical study of Italian political elections from Twitter data. *PloS one, 9(5*), e95809. DOI: 10.1371/journal.pone.0095809

Cameron, M. P., Barrett, P., & Stewardson, B. (2016). Can social media predict election results? Evidence from New Zealand*. Journal of Political Marketing, 15(4),* 416-432. DOI: 10.1080/15377857.2014.959690

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16(2),* 340-358. DOI: 10.1177/1461444813480466

Chang, J., Rosenn, I., Backstrom, L., & Marlow, C. (2010). ePluribus: Ethnicity on Social Networks*. ICWSM, 10,* 18-25.

Chen, B. (2009). Latent topic modelling of word co-occurence information for spoken document retrieval. DOI: 10.1109/ICASSP.2009.4960495

Chen, L., Wang, W., & Sheth, A. P. (2012, December). Are Twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican Presidential Primaries. *In International Conference on Social Informatics* (pp. 379-392). Springer, Berlin, Heidelberg.

Choy, M., Cheong, M. L., Laik, M. N., & Shung, K. P. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520.

Choy, M., Cheong, M., Laik, M. N., & Shung, K. P. (2012). US presidential election 2012 prediction using census corrected Twitter model. arXiv preprint arXiv:1211.0938.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010, December). Who is tweeting on Twitter: human, bot, or cyborg*?. In Proceedings of the 26th annual computer security applications conference* (pp. 21-30). ACM. DOI: 10.1145/1920261.1920265

Chung, J. E., & Mustafaraj, E. (2011, April). Can collective sentiment expressed on twitter predict political elections?. *In AAAI* (Vol. 11, pp. 1770-1771).

Clavel, C., & Callejas, Z. (2016). Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing, 7(1),* 74-93. DOI: 10.1109/TAFFC.2015.2444846

Couper, M. P. (2013, December). Is the sky falling? New technology, changing media, and the future of surveys. *In Survey Research Methods* (Vol. 7, No. 3, pp. 145-156). DOI: 10.18148/srm/2013.v7i3.5751

Davis, D. H. (2017, July). Is Twitter a Generalizable Public Sphere?: A Comparison of 2016 Presidential Campaign Issue Importance among General and Twitter Publics. *In Proceedings of the 8th International Conference on Social Media & Society* (p. 31). ACM.Chicago. DOI: 10.1145/3097286.3097317

Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PloS one, 11(1)*, e0145406. DOI: 10.1371/journal.pone.0145406

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one, 8(11),* e79449. DOI: 10.1371/journal.pone.0079449

Dimitrova, D. V., Kaid, L. L., Williams, A. P., & Trammell, K. D. (2005). War on the Web: The immediate news framing of Gulf War II. *Harvard International Journal of Press/Politics, 10(1),* 22-44. DOI: doi.org/10.1177/1081180X05275595

Dokoohaki, N., Zikou, F., Gillblad, D., & Matskin, M. (2015, August). Predicting swedish elections with twitter: A case for stochastic link structure analysis. *In Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference* on (pp. 1269-1276). IEEE. DOI: 10.1145/2808797.2808915

Flemming, G., & Sonner, M. (1999, May). Can Internet polling work? Strategies for conducting public opinion surveys online. *In annual meeting of the American Association for Public Opinion Research*, St. Petersburg Beach, FL.

Gamson, W. A. (1989). News as framing: Comments on Graber. *American behavioral scientist, 33(2),* 157-161. DOI: 10.1177/0002764289033002006

Gayo-Avello, D. (2011). Don't turn social media into another'Literary Digest'poll. *Communications of the ACM, 54(10),* 121-128. DOI: 10.1145/2001269.2001297

Gayo-Avello, D. (2012a). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data. arXiv preprint arXiv:1204.6441.

Gayo-Avello, D. (2012b). No, you cannot predict elections with Twitter. *IEEE Internet Computing, 16(6),* 91-94. DOI: 10.1109/MIC.2012.137

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review, 31(6),* 649-679. DOI: 10.1177/0894439313493979

Ghosh, S., Zafar, M. B., Bhattacharya, P., Sharma, N., Ganguly, N., & Gummadi, K. (2013, October). On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. *In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 1739-1744). ACM. DOI: 10.1145/2505515.2505615

Goldstein, P., & Rainey, J. (2010). The 2010 elections: Twitter isn't a very reliable prediction tool. *Retrieved January*, 10, 2012.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on T witter. *Journal of the Association for Information Science and Technology, 67(1),* 232-238. DOI: 10.1002/asi.23456

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social science computer review, 30(2),* 229-234. DOI: 0.1177/0894439311404119

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research, 23(5),* 544-559. DOI: 10.1108/IntR-06-2012-0114

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016, April). Inferring gender from names on the web: A comparative evaluation of gender detection methods. *In Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 53-54). International World Wide Web Conferences Steering Committee. DOI: 10.1145/2872518.2889385

Khatua, A., Khatua, A., Ghosh, K., & Chaki, N. (2015, January). Can# twitter_trends predict election results? Evidence from 2014 indian general election. *In System Sciences (HICSS),* 2015 48th Hawaii International Conference on (pp. 1676-1685). IEEE. DOI: 10.1109/HICSS.2015.202

Krstajić, M., Mansmann, F., Stoffel, A., Atkinson, M., & Keim, D. A. (2010, March). Processing online news streams for large-scale semantic analysis. *In Data Engineering Workshops (ICDEW),* 2010 IEEE 26th International Conference on (pp. 215-220). IEEE. DOI: 10.1109/ICDEW.2010.5452710

Li, H., Jou, B., Ellis, J. G., Morozoff, D., & Chang, S. F. (2013, October). News rover: exploring topical structures and serendipity in heterogeneous multimedia news. *In Proceedings of the 21st ACM international conference on Multimedia* (pp. 449-450). ACM. DOI: 10.1145/2502081.2502263

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the predictability of the US elections through search volume activity.

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets*. People, 1*(3,759.710), 3-759.

Marchetti-Bowick, M., & Chambers, N. (2012, April). Learning for microblogs with distant supervision: Political forecasting with twitter*. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 603-612). Association for Computational Linguistics.

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections*. In Privacy, Security, Risk and Trust (PASSAT)* and *2011 IEEE Third International Conference on Social Computing (SocialCom),* 2011 *IEEE Third International Conference on* (pp. 165-171). IEEE. DOI: 10.1109/PASSAT/SocialCom.2011.98

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM, 11(5th)*, 25.

Morstatter, F., Dani, H., Sampson, J., & Liu, H. (2016, April). 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.DOI: 10.1145/2872518.2889372

Morstatter, F., Pfeffer, J., & Liu, H. (2014, April). When is it biased?: assessing the representativeness of twitter's streaming API. *In Proceedings of the 23rd International Conference on World Wide Web* (pp. 555-556). ACM.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013, July). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., ... & Harwood, P. (2014). Social media in public opinion research: executive summary of the Aapor task force on emerging technologies in public opinion research. *Public Opinion Quarterly, 78(4),* 788-794. DOI: 10.1093/poq/nfu053

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm, 11*(122-129), 1-2.

O'Leary, D. E. (2015). Twitter mining for discovery, prediction and causality: Applications and methodologies. *Intelligent Systems in Accounting, Finance and Management, 22(3),* 227-247. DOI: 10.1002/isaf.1376

Papadopoulos, S., Corney, D., & Aiello, L. M. (2014, April). SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media. In SNOW-DC@ *WWW* (pp. 1-8).

Petrovic, S., Osborne, M., & Lavrenko, V. (2011). Rt to win! predicting message propagation in twitter. *ICWSM, 11,* 586-589.

Phillips, L., Dowling, C., Shaffer, K., Hodas, N., & Volkova, S. (2017). Using social media to predict the future: a systematic literature review. arXiv preprint arXiv:1706.06134.

Pimenta, F., Obradovic, D., & Dengel, A. (2013, September). A comparative study of social media prediction potential in the 2012 us republican presidential preelections. *In Cloud and Green Computing (CGC),* 2013 Third International Conference on (pp. 226-232). IEEE. DOI: 10.1109/CGC.2013.43

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. *ICWSM, 11,* 297-304.

Reyes, A., & Rosso, P. (2014). On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems, 40(3),* 595-614.

Roeder, O., Mehta, D., & Wezerek, G. (2017, Oct 24). The worst Tweeter in politics isn't Trump. FiveThirtyEight Retrieved from https://fivethirtyeight.com/features/the-worst-tweeter-in-politics-isnt-trump/

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346(6213),* 1063-1064. DOI: 10.1126/science.346.6213.1063

Sang, E. T. K., & Bos, J. (2012, April). Predicting the 2011 dutch senate election results with twitter. *In Proceedings of the workshop on semantic analysis in social media* (pp. 53-60). Association for Computational Linguistics.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public opinion quarterly, 80(1),* 180-211. DOI: 10.1093/poq/nfv048

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research, 23(5),* 528-543. DOI: 10.1108/IntR-06-2013-0115

Scott Keeter, Kyley McGeeney, Ruth Igielnik, Andrew Mercer, Nancy Mathiowetz (2015, May 13), *From Telephone to the Web: The Challenge of Mode of Interview Effects in Public Opinion Polls.* Retrieved from http://www.pewresearch.org/2015/05/13/from-telephone-to-the-web-the-challenge-of-mode-of-interview-effects-in-public-opinion-polls/

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of communication, 50(2),* 93-109. DOI: 10.1111/j.1460-2466.2000.tb02843.x

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19(1),* 22-36. DOI: 10.1145/3137597.3137600

Sjøvaag, H., & Stavelin, E. (2012). Web media and the quantitative content analysis: Methodological challenges in measuring online news content. *Convergence, 18(2),* 215-229. DOI: 10.1177/1354856511429641

Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., & Jiang, J. (2012, January). Tweets and votes: A study of the 2011 singapore general election. *In System Science (HICSS),* 2012 45th Hawaii International Conference on (pp. 2583-2591). IEEE. DOI: 10.1109/HICSS.2012.607

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics, 118,* 26-40. DOI: 10.1016/j.jpubeco.2014.04.010

Stieglitz, S., & Dang-Xuan, L. (2012, January). Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. *In System Science (HICSS),* 2012 45th Hawaii International Conference on (pp. 3500-3509). IEEE. DOI: 10.1109/HICSS.2012.476

Tian, Y., & Stewart, C. M. (2005). Framing the SARS crisis: A computer-assisted text analysis of CNN and BBC online news reports of SARS. *Asian Journal of Communication, 15(3),* 289-301. DOI: 10.1080/01292980500261605

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*. ICWSM, 14,* 505-514.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment*. Icwsm, 10(1),* 178-185.

Wang, C., Zhang, M., Ru, L., & Ma, S. (2008, October). Automatic online news topic ranking using media focus and user attention based on aging theory. *In Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1033-1042). ACM. DOI: 10.1145/1458082.1458219

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls*. International Journal of Forecasting, 31(3),* 980-991. DOI: 10.1016/j.ijforecast.2014.06.001

Weller, K. (2015). Accepting the challenges of social media research. *Online Information Review, 39(3),* 281-289. DOI: 10.1108/OIR-03-2015-0069

Williams, C. B., & Gulati, G. (2008). The political impact of Facebook: Evidence from the 2006 midterm elections and 2008 nomination contest. *Politics and Technology Review, 1(1),* 11-24.

Yu, S., & Kak, S. (2012). A survey of prediction using social media. arXiv preprint arXiv:1203.1647.

Zhang, W., Ram, S., Burkart, M., & Pengetnze, Y. (2016, April). Extracting signals from social media for chronic disease surveillance. *In Proceedings of the 6th International Conference on Digital Health Conference* (pp. 79-83). ACM. DOI: 10.1145/2896338.2897728

Biographical Notes

Jin-ah KWAK is a Ph.D. candidate at KAIST in Daejeon, South Korea. Big data analysis is one of her major research interests.

She can be reached at KAIST, 291 Daehak-ro, Eoeun-dong, Yuseong-gu, South Korea or by email at jinah.kwak@kaist.ac.kr.

Sung Kyum CHO is a professor in the Department of Communication at Chungnam National University and director of the Center for Asian Public Opinion Research & Collaboration Initiative (CAPORCI). He was the first president of the Asian Network for Public Opinion Research (ANPOR). He has also been president of the Korean Association for Survey Research (KASR) and the Korean Society for Journalism and Communication Studies (KSJCS). He is part of the team that conducts the Korean Academic Multimode Open Survey (KAMOS). He is also an associate editor for and publisher of the Asian Journal for Public Opinion Research (AJPOR).

He can be reached at Chungnam National University, Department of Communication 99, Daehak-ro, Yuseong-gu, Daejeon, 305-764 or by email at skcho@cnu.ac.kr.

Date of Submission: 2018-08-19
Date of the Review Results: 2018-08-25
Date of the Decision: 2018-08-28