Hybrid method in revealing facts behind texts: A combination of text mining and qualitative approach

(Sub-theme: Exploring New Research Methods for Public Administration: beyond positivism)

Ujang Fahmi[1], Canggih Puspo Wibowo[2], Faturahman Yudanto[3]
Department of Public Policy and Management, Universitas Gadjah Mada, Sagasitas Journal & Research Center, Yogyakarta, Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada
ujang.fahmi@mail.ugm.ac.id[1], canggih.p.w@ieee.org[2], f.yudanto@mail.ugm.ac.id[3]

Keywords: analysis method, text mining, social media, soft-data, qualitative, quantitative

Abstract

Social media has become one of the primary sources of data which available for policy analysts and policymakers. As the evidence, active Twitter users are sending 500 million tweets per day containing thoughts, opinions, pictures, and other information. Social media offers new challenges related to how the data is acquired and how to analyze it. Unfortunately, the state-of-the-art methods in text mining are still unable to interpret texts fully. Thus, in social media analysis, we can only make a conclusion based on the insight into an event. Therefore, we propose a hybrid method that combines text mining and qualitative methods for analyzing social media data. This research was composed based on a review of studies and experimental results on the data taken from the Twitter. The results show that both techniques can complement each other and give in-depth analysis of the data. Furthermore, the results can be employed to observe social media data in a faster, cheaper, and more precise way. More importantly, the results of this study serve as a basis for further development of a method to reveal the facts behind texts that obtained from social media.

Introduction

The availability of social media as a data source for both social researchers and policy-makers not only provides an opportunity to get more data faster, but it also presents challenges. Some researchers focus on building techniques for acquiring the data, while some others were presenting approaches to analyze it. By 2017, there are 330 million active Twitter users sending 500 million tweets per day containing thoughts, opinions, pictures, and other information (Statista, 2017). From that huge number of data obtained from the Twitter, we can study many things, such as: (1) public sentiment related to a policy (Ceron & Negri, 2015); (2) Public policy making preferences and processes (Koltsova & Koltcov, 2013; Severo, Giraud, & Pecout, 2016); (3) Community mapping (Batorski & Grzywińska, 2017); (4) Communication strategy (Floreddu & Cabiddu, 2016); (5) Community collective campaigns and actions (Ceron, 2017; van den Heerik, van Hooijdonk, Burgers, & Steen, 2017); (6) Promotion of transparency, and accountability of public organizations (Bertot, Jaeger, & Grimes, 2012); (7) Mapping and knowing public opinion (Jaeger, Bertot, & Shilton, 2012; Koltsova & Koltcov, 2013). Therefore, Leavey (2013, p. 5) says that social media presents a growing body of

evidence that can inform social and economic policy. It has value for government, the policy community, and public service delivery organization.

However, the amount of research and evidence that can be obtained from social media data often collided with some issues accompanying it (Bright, Margetts, Hale, & Yasseri, 2014). Some of these concerns include the use of application programming interfaces (APIs) to obtain data providing only limited data available with the maximum period of seven days (Vick, Soporan, Lewis, & Zurn, 2012, p. 30). Limitations on the amount and duration of data then raise issues of trust and representation of data (Leavey, 2013, p. 22; Marcellino, Smith, Paul, & Skrabala, 2017, p. 28). Since the unstructured data analysis of social media is still in the development stage, the method used often raises questions about its validity. The validation issue arises because large amounts of data are analyzed using only a quantitative approach (Marwick, 2014, p. 118).

Given the considerable potential and limitations of current methods that can be used to gain insight from large amounts of data, we propose a hybrid method as a bridge unto a reliable method for analyzing "soft data." The term "soft data" refers to data from social media that is open to the public and is bottom-up (Severo, Feredj, & Romele, 2016, p. 43). This research aims to answer the following questions: (1) What and why hybrid method? (2) How hybrid methods used to reinforce the results of the analysis? (3) How is the application in the case study? And (4) what is the prospect?

This article is written in the following order: First, we explain our proposed hybrid method by providing the fundamental concepts of the approach. In this part, we also discuss how potential is the hybrid method to strengthen the results of soft data analysis. Second, we show the results of the hybrid method application in a case study. Third, we elaborate the advantages of the hybrid method. Finally, we present conclusion by addressing the prospect of the hybrid method to be used in analyzing soft data.

Hybrid method

Text mining method has many benefits for both analysts and policy-makers to get and process soft data quickly (Bicquelet & Weale, 2011). As for addressing and dealing with the weaknesses associated with the text mining approach, Bicquelet and Weale (2011) suggest combining quantitative and qualitative analysis to reduce both technical and ethical shortcomings. Therefore, it needs a combination of text mining that refers to the process of extracting information and knowledge from unstructured text (Hotho, Nürnberger, & Paaß, 2005) with other methods or approaches in analyzing soft data.

A hybrid method we proposed focuses on revealing the facts behind texts. It combines the fast-quantitative approach in text mining and the in-depth analysis of qualitative approach. With the combination of both, we assume that the underlying information can be gathered in more detail and accurate than only using either. Figure 1 shows our proposed hybrid method to explore the underlying information from soft data. We compose three steps of procedure representing a mixed way between text mining and qualitative approach.

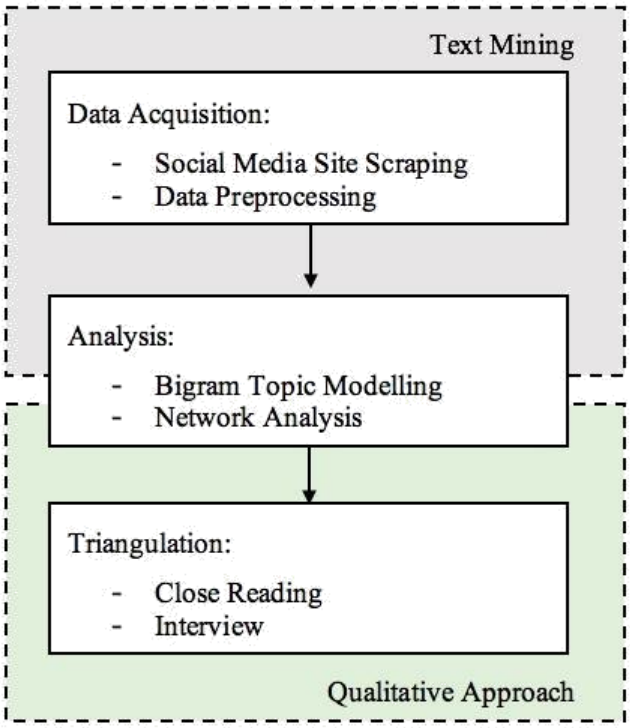

Figure 1. Hybrid method procedure

Data Acquisition

Most soft data researchers use application programming interface (API) to extract data from web service in social media sites (Vick et al., 2012, p. 30). However, API provided by the web service offer some limitations, especially in a free scheme. Typically, free API is constrained by the amount of data collected, the number of requests per day, and what kind of data provided, hence making it difficult for soft data researchers to acquire adequate data. The social media owner only offers API with extensive capabilities to those who are willing to pay for the premium scheme. The limitations in obtaining data make the information extracted from the soft data may not reflect the reality. Our approach to overcome this condition is to scrap the social media website and search for particular data directly from its HTML sources. This method is indeed slower and consuming more computational resources than using API. Nevertheless, it will not be bounded by any limitations.

The data obtained from scraping step are unstructured. Thus, before going further analyzing, data preprocessing is needed (Angiani et al., 2016; Uysal & Gunal, 2014). Preprocessing includes tokenization, stemming, characters and words removal, such as digits, stop words, and some non-significant words. In this work, removing words which are not contributing to the topic discussed is an essential step. We consider such words as noise. The non-significant words are selected manually after observing the data. We do the same preprocessing method as Fahmi et al. (2017).

Analysis

Some researchers have used various methods in the study of public opinion mapping. Sokolova et al. (2016) and Fahmi et al. (2017) use topic modeling to reveal the what is discussed in social media. Etling et al. (2014) and Shapiro and Hemphill (2017) use content analysis to find out the similarities and differences of various media content such as TV,

Mainstream Media, and Blogs. While Crump (2011) uses content and network analysis to see the impact of efforts to increase public confidence through social media, content analysis principle is also applied by Dredze (2012) to develop Ailment Topic Aspect Model (ATAM), a probabilistic graphical model that is used to explore tweets containing messages about health. More specific, topic modeling method used by Koltsova and Koltcov (2013) to find out public agenda through citizens blog posts.

In this work, to reveal facts behind texts, we carry out two calculations, which are topic modeling and network analysis. Topic modeling here can be employed to find the central theme of a large and unstructured text document (Blei, 2010) while network analysis is done to find out the influencers, i.e., people who influence others in particular topics. Based on our preliminary studies, we realize that in social media text there are many local contexts in the form of two-words combination. Therefore, we consider using bigram representation to emphasize the local meaning of information, instead of single-independent words. Some studies also stated the significance of bigram representation in text mining, such as in Tan et al. (2002) and Dunning (1993). Afterwards, we use Latent Dirichlet Allocation (LDA) (Blei, 2010) on the bigram representation to model the topics.

An in-depth analysis based on qualitative approach is conducted for the relation between user accounts who posted topics. Here, we take the user accounts and create a network representation based on their proximity to others. Proximity here is calculated by how much a user account mention other in the post. With this proximity measurement, we construct the network. Previous researchers have been using a network representation to observe the group detection, group profiling, and recommendation system (Barbier, 2011). It has an advantage in the use of mathematical graph theory (Butts, 2009) to enhance the analysis.

The goal of the network representation is to reveal the influencers, which we consider as the central figures of the topic discussion. In the graph theory, the influencers can be observed by computing the centralities of the graph. There are some centralities proposed by researchers. However, we only employ two kinds of centralities based on local and global approach, which are the degree (Nieminen, 1973) and betweenness centrality (Freeman, 1977).

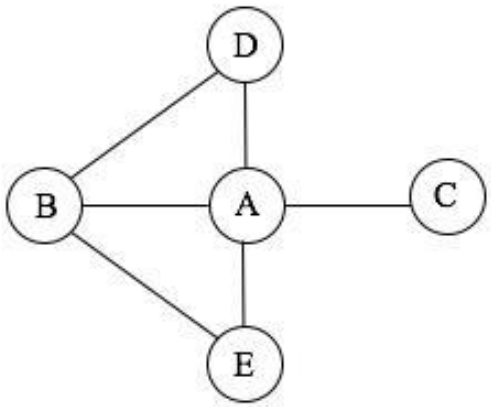

Figure 2. Network illustration

Degree centrality, which considered as a local approach centrality, measures the influencers based on the number of neighboring people in the network. For example, in Figure 2, node A has degree centrality of 4 since it is connected to four nodes, while E has only the

degree of 2. On the other hand, betweenness centrality uses a global approach, that is calculating how often a path lies on the influencers. From Figure 2, we can deduct that node A has a significant value of betweenness because all path connecting node C have to pass node A. Furthermore, with the combination of local and global approaches, we extract a more in-depth analysis of the influencers. The two centralities have been used in some works including (Mergel, 2017, p. 4) and (Chung & Zeng, 2016, p. 1595).

Triangulation

In the final step, triangulation is an essential process in the study because social media especially Twitter is social media with a high level of anonymity (Chen, 2013, p. 75; Kaplan & Haenlein, 2010, p. 62). Here, we follow up with a qualitative approach as suggested by Marwick (2014), i.e., close reading and virtual ethnography. We read the data carefully and interview specific account owners who have become influencers based on the network analysis.

By combining these approaches, it is easier for us to get and process data to support research, we also get more information that until now difficult to be obtained. Using a hybrid approach, the results of the study conducted are even more profoundly and provided a real knowledge of the growing public opinion in the social media. So, although it still requires manual efforts, we consider this method is appropriate to be used to get and analyze data from the social media.

Case study on #jogjaoradidol hashtag

Our hybrid method was applied to analyze Twitter data which having #jogjaoradidol hashtag. We chose this hashtag because it has been a viral hashtag in Yogyakarta Special Region since October 2013. The hashtag #jogjaoradidol was initially used as a mural painting on one of the historical building corners in Yogyakarta. Later, it also has been found in various other media such as songs, movie titles, and even social movements with the same name, namely "jogja ora didol" movement. So, the existence of the slogan "jogja ora didol" which dissemination is much helped by the social media especially Twitter (Fanggidae, 2016, p. 177) in the community can be categorized as a phenomenon. Here we would like to know the answer to the central question: "what is discussed by users on the Twitter by using #jogjaoradidol hashtag?"

We collected Twitter tweets with #jogjaoradidol hashtag from 7 October 2013 until 31 August 2017 consisting 16,456 tweets posted by 7,708 different Twitter accounts. The result of topic modeling analysis is shown in Table 1. With the long period of data, it provides some indicator of the development of the topics covered. For example, in 2013 that started from tweets in October we can find a topic about a birthday (Table 1), where the birthday refers to the anniversary of the city of Yogyakarta which falls in the same month. While from topic 2 in 2013, we get a few topics like “beton/concrete,” “pohon/tree,” “mall,” “hotel,” “kota_seni/art city” and “satpol_pp/public order enforcers."

Some of the topics mentioned above are related to the refusal and protests of the residents associated with the implementation of development policy in the city of Yogyakarta. Where people feel that development should be done by "planting trees" instead of "concrete." Here through tweets, we can also identify that the word "hotel" and "mall" are part of the trigger for tweets that use hashtag #jogjaoradidol. Overall, Table 1 lists some consistent topics

emerging each year such as birthdays and some new things that come up like the topic "warga_gadingan/citizen of Gadingan," "warga_karangwuni/citizens of Karangwuni." Several topics which emerged at the beginning also experienced developments such as "ijin_hotel/permit_hotel" which all we can only find "mall_hotel" and the topic that emerged in the year 2017 is "jalan_tol/toll road."

Table 1. Topic modeling per year

Category	2013 (07 October – 31 Desember 2013)		2014 (01 Januari– 31 Desember 2014)		2015 (01 Januari– 31 Desember 2015)		2016 (01 Januari– 31 Desember 2016)		2017 (01 Januari– 31 August 2017)	
Total tweets	**5365**		**5729**		**3655**		**1304**		**403**	
Topic 1	0.013*	puncak_hut	0.005*	tetap_istimewa	0.005*	warga_gadingan	0.003*	izin_hotel	0.003*	jalan_tol
	0.005*	video_dukungan	0.002*	pembangunan_hotel	0.004*	warga_karangwuni	0.002*	sultan_merasa	0.002*	perintahkan_penghentian
	0.003*	pas_puncak	0.002*	no_placard	0.003*	ulang_tahun	0.002*	soal_izin	0.002*	proses_pembangunan
	0.003*	dukungan_sid	0.002*	metro_tv	0.003*	tetap_istimewa	0.002*	merasa_ditipu	0.002*	balirejo_atas
	0.002*	tamu_spesial	0.002*	semoga_tetap	0.003*	bawah_tanah	0.002*	hotel_sultan	0.002*	penghentian_proses
Topic 2	0.004*	bukan_beton	0.013*	alunalun_utara	0.003*	pembangunan_hotel	0.008*	gawe_kebon	0.004*	satpol_pp
	0.003*	tanam_pohon	0.013*	siap_berburu	0.002*	hotel_baru	0.007*	nandur_beton	0.004*	warga_balirejo
	0.003*	mall_hotel	0.005*	ulang_tahun	0.002*	marak_campaign	0.002*	pembangunan_hotel	0.003*	tolak_pembangunan
	0.002*	kota_seni	0.003*	posting_sebar	0.001*	hotel_mall	0.001*	hari_ini	0.003*	tetap_istimewa
	0.002*	satpol_pp	0.002*	cara_bijak	0.001*	kasus_penghancuran	0.001*	penguasa_rakus	0.002*	aset_pribadi

Source: obtained from primary data

* The probability of a word that appears on the topic

Table 2. Topic Structures of #jogjaoradidol tweets from 07 October, 2013 to 31 August, 2017

No	Issue	Topic and Year
1	Area (5 topics - 10 %)	alunalun_utara (2014), warga_gadingan (2015), warga_karangwuni (2015), balirejo_atas (2017), warga_balirejo (2017)
2	Birthday (4 topics - 8%)	puncak_hut (2013), pas_puncak (2013), ulang_tahun (2014), ulang_tahun (2015)
3	Area Identity (5 topics - 10%)	kota_seni (2013), semoga_tetap (2014), tetap_istimewa (2014), tetap_istimewa (2015), tetap_istimewa (2017)
4	Refusal Construction of hotels and malls (18 topics - 36%)	bukan_beton (2013), tanam_pohon (2013), mall_hotel (2013), pembangunan_hotel (2014), pembangunan_hotel (2015), hotel_baru (2015), hotel_mall (2015), kasus_penghancuran (2015), izin_hotel (2016), soal_izin (2016), gawe_kebon (2016), nandur_beton (2016), pembangunan_hotel (2016), jalan_tol (2017), perintahkan_penghentian (2017), proses_pembangunan (2017), penghentian_proses (2017), tolak_pembangunan (2017)
5	Government/individual related to power (7 topics - 14%)	satpol_pp (2013, sultan_merasa (2016), merasa_ditipu (2016), hotel_sultan (2016), penguasa_rakus (2016), satpol_pp (2017), aset_pribadi (2017)
6	Support (4 topics - 8%)	video_dukungan (2013), dukungan_sid (2013), posting_sebar (2014), marak_campaign (2015)
7	Uninterpreted topics (7 topics - 14%)	tamu_spesial (2013), metro_tv (2013), no_placard (2013), siap_berburu (2014), cara_bijak (2014), bawah_tanah (2015), hari_ini (2016)

Source: obtained from primary data

Table 2 shows the topic modeling result of the data gathered. To explore more detail about these topics, we conducted a qualitative study based on the network analysis. The result of network analysis on tweets with #jogjaoradidol hashtag is shown in Figure 3. After performing a calculation on the centrality, we concluded several accounts which are classified as the influencers, visualized in Figure 3. In this work, we only chose account @dodoputrabangsa and @joeyakarta for the qualitative method approach. Among all influencers shown in Figure 3, only @dodoputrabangsa and @joeyakarta who posted many tweets in the period of data taken, which are 475 and 188 respectively.

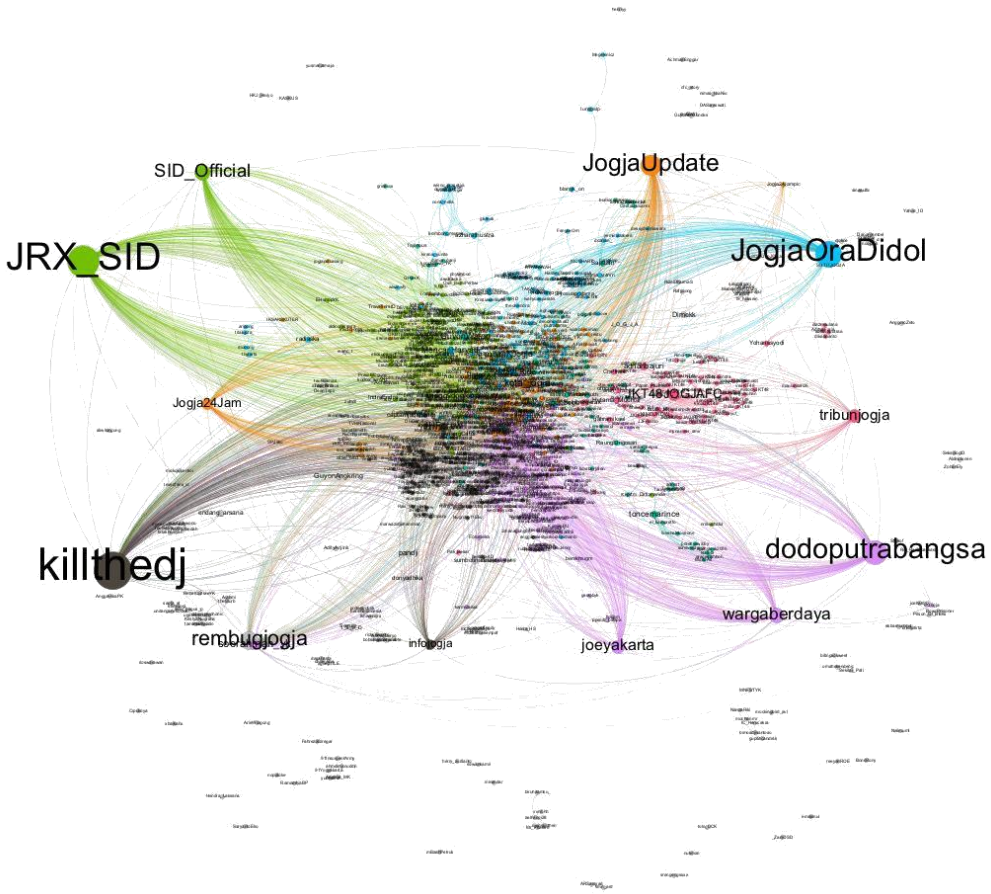

Figure 3. The network of #jogjaoradidol with degree centrality more than 4, showing the influencers based on the betweenness centrality. Text size shows the betweenness centrality value. The higher the centrality value, the bigger text size. Line width shows the proximity of users. The thick lines mean proximate. Line color shows the group, calculated from the topics.

From the network analysis, we also calculated the group of users based on the topic they discussed. As can be seen in Figure 3, there are several groups identified by the line colors. It is essential to know the groups for further qualitative observation. For example, @dodoputrabangsa, @joeyakarta, and @wargaberdaya are in the same category. Thus, we can assume that they are in the same class in discussing the #jogjaoradidol hashtag. This is also one reason why we chose to select @dodoputrabangsa and @joeyakarta for further analysis.

Close reading and interviewing the influencers, both @dodoputrabangsa and @joeyakarta, give us in-depth information related to the topic presented in Table 2. The majority of tweets posted are used to discuss the rejection of hotel constructions. For example, from the tweets posted by @dodoputrabangsa and @joeyakarta, we can see that the "kasus_penghancuran/case of destruction" topic represents the case of building a hotel by demolishing a historical building. From the tweets posted by the two accounts, we can obtain information about the case and also get the viewpoint of the person who posts the tweet. Hence, we can do a more in-depth study of the particular topic.

Through this case study, we can confidently say that tweets using #jogjaoradidol hashtag are used by the people to express their disagreement with the regional development process, especially for malls and hotels in Yogyakarta. The social movement of “jogja ora didol”also receive public support coming from various circles, ranging from ordinary people, artists, to politicians. The results of this case study confirm the opinion of Fanggidae (2016, p. 177) which said that the "jogja ora didol" social movement is supported by the existence of social media, especially Twitter.

Advantages of hybrid method

Based on literature studies and experiments conducted in the case studies discussed in the previous section, the methods we offer have several benefits to answer the challenge of revealing the facts behind text taken from social media data. The advantages are summarized as follows

Retrieve more data

Our proposed method ensures the data to be obtained according to what we want. In the case study, we managed to get the data in the form of tweets for an extended period. Here we can observe the first tweet containing a particular hashtag. Analyzing the context of the tweet is important. In future research, this approach in data retrieval allows researchers to conduct longitudinal data analysis of events that have occurred and become a topic of discussion in social media. Besides, a more significant amount of data may also increase confidence in the results of the analysis performed.

More readable topic

Typically, topic modeling is conducted with only one word (unigram), but in the case tweets coming from Indonesia, it is less readable due to some mixed languages. Our hybrid method anticipates this by providing a comprehensive preprocessing step. Also, by using bigram representation, many contexts involving a pair of words can be extracted.

Able to observe the influencers and groups

With the aim to deepen and gain the facts of the growing discourse on Twitter, the use of network analysis here proved to be helpful. The influencers in the network are shown to have a high number of centrality. It becomes the primary source for the qualitative approach. Additionally, we also can analyze the groups of users sharing the same topic of discussion.

Reveal the context of the text

Interviews and outcomes can be used to ascertain existing communes, as in the case of ensuring the existence of a community of "powerless citizens." Interviews can be conducted via direct conversations on Twitter. This approach produces a very particular and constrained style of interview, due to the 140-character limit (Marwick, 2014, p. 110). While using the close reading method, in which texts are read paying rigorous attention to individual words,

syntax, and diction (Marwick, 2014, p. 118), we can understand the context that the Twitter users are trying to construct with their tweets.

Conclusion and future research direction

Qualitative methods, such as interviews, ethnographic observations, and content analysis can provide us with the data beyond descriptions. Through this paper, we conclude that text mining and qualitative approach are complements to each other. Where from the case study, topic modeling with bigram and network analysis can be used as the basis of the qualitative data collection. It also minimizes the impact of non-normalized text and word stemming in the preprocessing step. While the qualitative data obtained can then be used to confirm and deepen the description of the findings. The results of close reading and interviews with the user account owner enabled us to know the context of a topic discussed on Twitter and knew the real community. More importantly, using the hybrid method we propose here, an analysis of a phenomenon that emerges in social media can be done thoroughly and enables the process of triangulating data. The triangulation process itself can be used to answer the issue of representation which until now is still a big issue in the utilization of soft data. Thus, with the triangulation process, the level of confidence in the results of research can be increased, and the facts behind the text can be revealed. For that reason, in the future, the combined method needs to be further refined with different case studies and research questions.

References

Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., & Manicardi, S. (2016). A comparison between preprocessing techniques for sentiment analysis in Twitter. *CEUR Workshop Proceedings*, *1748*, 1–11. https://doi.org/10.1007/978-3-319-67008-9_31

Barbier, G. (2011). Data mining in social media. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 327–352). New York: Springer. https://doi.org/10.1007/978-1-4419-8462-3

Batorski, D., & Grzywińska, I. (2017). Three dimensions of the public sphere on Facebook. *Information, Communication & Society*, 1–19. https://doi.org/10.1080/1369118X.2017.1281329

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2012). Promoting transparency and accountability through ICTs, social media, and collaborative e-government. *Transforming Government: People, Process and Policy*, *8*(2), 283–308. https://doi.org/doi:10.1108/TG-08-2013-0026

Bicquelet, A., & Weale, A. (2011). Coping with the Cornucopia: Can Text Mining Help Handle the Data Deluge in Public Policy Analysis? *Policy & Internet*, *3*(4), 1–21. https://doi.org/10.2202/1944-2866.1096

Blei, D. M. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*(6), 77–84. https://doi.org/10.1109/MSP.2010.938079

Bright, J., Margetts, H., Hale, S., & Yasseri, T. (2014). *The Use of Social Media for Research and Analysis: A Feasibility Study*. London. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/387591/use-of-social-media-for-research-and-analysis.pdf

Bruns, A., & Stieglitz, S. (2014). Metrics for Understanding Communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahart, & C. Puschmann (Eds.), *Twitter and Society* (pp. 69–82). New York: Peter Lang. Retrieved from http://www.peterlang.com/index.cfm?event=cmp.ccc.seitenstruktur.detailseiten&seitentyp=produkt&pk=71177&cid=5&concordeid=312169

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science*, *325*(5939), 414–416. https://doi.org/10.1126/science.1171022

Ceron, A. (2017). Social Media, Collective Action and Public Policy. In *Social Media and Political Accountability* (pp. 133–156). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-52627-0_7

Ceron, A., & Negri, F. (2015). Public policy and social media: How sentiment analysis can support policy-makers across the policy cycle. *Rivista Italiana Di Politiche Pubbliche*, *10*(3), 309–338. https://doi.org/10.1483/81600

Chen, P. J. (2013). Social media. In *Australian Politics in a Digital Age*. Canberra: ANU E Press. Retrieved from http://www.jstor.org/stable/j.ctt2jbkkn.11

Chung, W., & Zeng, D. (2016). Social-media-based public policy informatics: Sentiment and network analyses of U.S. Immigration and border security. *Journal of the Association for Information Science and Technology*, *67*(7), 1588–1606. https://doi.org/10.1002/asi.23449

Crump, J. (2011). What Are the Police Doing on Twitter? Social Media, the Police and the Public. *Policy & Internet*, *3*(4), 1–27. https://doi.org/10.2202/1944-2866.1130

Dredze, M. (2012). How social media will change public health. *IEEE Intelligent Systems*, *27*(4), 81–84. https://doi.org/10.1109/MIS.2012.76

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74. Retrieved from http://portal.acm.org/citation.cfm?id=972454

Etling, B., Roberts, H., & Faris, R. (2014). *Blogs as an Alternative Public Sphere: The Role of Blogs, Mainstream Media, and TV in Russia's Media Ecology*. *The Berkman Center for Internet & Society Research Publication Series* (Vol. 8). https://doi.org/10.2139/ssrn.2427932

Fahmi, U., & Wibowo, C. P. (2017). Online public sphere: a new dimension in the public policy-making process (Case study of Twitter utilization in Indonesia). In M. Taufiq (Ed.), *International Seminar: Reconstructing Public Administration Reform to Build World Class Government* (pp. 591–605). Jakarta: Lembaga Administrasi Negara.

Fahmi, U., Wibowo, C. P., & Yudanto, F. (2017). Mapping the principal-agent relation in social media: A data mining approach. In *International Seminar on Social and Political Science* (pp. 1–16). Yogyakarta: Fisipol, UGM.

Fanggidae, I. G. (2016). *Implikasi Gerakan "Jogja Ora Didol" Terhadap Penetapan Agenda Kebijakan di Kota Yogyakarta*. Universitas Gadjah Mada.

Floreddu, P. B., & Cabiddu, F. (2016). Social media communication strategies. *Journal of Services Marketing*, *34*(6), 754–776. https://doi.org/10.1108/JSM-01-2015-0036

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35. https://doi.org/10.2307/3033543

Gaffney, D., & Puschmann, C. (2014). Data Collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahart, & C. Puschmann (Eds.), *Twitter and Society* (pp. 55–67). New York: Peter Lang.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *Journal for Computational Linguistics and Language Technology*, *20*, 19–62. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Jaeger, P. T., Bertot, J. C., & Shilton, K. (2012). Information Policy and Social Media: Framing Government—Citizen Web 2.0 Interactions. In *Web 2.0 Technologies and Democratic Governance -Political, Policy and Management Implications* (pp. 11–26). https://doi.org/10.1007/978-1-4614-1448-3

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Koltsova, O., & Koltcov, S. (2013). Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal. *Policy and Internet*, *5*(2), 207–227. https://doi.org/10.1002/1944-2866.POI331

Leavey, J. (2013). *Social media and public policy: what is the evidence?* London. Retrieved from http://www.alliance4usefulevidence.org/publication/social-media/

Marcellino, W., Smith, M., Paul, C., & Skrabala, L. (2017). *Monitoring Social Media: Lessons for Future Department of Defense Social Media Analysis in Support of Information Operations*. *RAND Corporation*. Santa Monica: RAND Corporations. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=27460084&site=bsi-live

Marwick, A. E. (2014). Ethnographic and Qualitative Research on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahart, & C. Puschmann (Eds.), *Twitter and Society* (pp. 109–122). New York: Peter Lang.

Mergel, I. (2017). Building Holistic Evidence for Social Media Impact. *Public Administration Review*, *77*(4), 489–495. https://doi.org/10.1111/puar.12780

Nieminen, U. J. (1973). On the centrality in a directed graph. *Social Science Research*, *2*(4), 371–378. https://doi.org/10.1016/0049-089X(73)90010-0

Severo, M., Feredj, A., & Romele, A. (2016). Soft Data and Public Policy: Can Social Media Offer Alternatives to Official Statistics in Urban Policymaking? *Policy and Internet*, *8*(3), 354–372. https://doi.org/10.1002/poi3.127

Severo, M., Giraud, T., & Pecout, H. (2016). *Twitter data for urban policy making: an analysis on four European cities*. Luxembourg.

Shapiro, M. A., & Hemphill, L. (2017). Politicians and the Policy Agenda: Does Use of Twitter by the U.S. Congress Direct New York Times Content? *Policy & Internet*, *9*(1), 109–132. https://doi.org/10.1002/poi3.120

Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., … Sambuli, N. (2016). Topic Modelling and Event Identification from Twitter Textual Data, (August), 17. Retrieved from http://arxiv.org/abs/1608.02519

Statista. (2017). Twitter MAU worldwide 2017. Retrieved November 5, 2017, from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Tan, C. M., Wang, Y. F., & Lee, C. Do. (2002). The use of bigrams to enhance text categorization. *Information Processing and Management*, *38*(4), 529–546. https://doi.org/10.1016/S0306-4573(01)00045-0

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

van den Heerik, R. A. M., van Hooijdonk, C. M. J., Burgers, C., & Steen, G. J. (2017). "Smoking Is Sóóó ... Sandals and White Socks": Co-Creation of a Dutch Anti-Smoking Campaign to Change Social Norms. *Health Communication*, *32*(5), 621–628. https://doi.org/10.1080/10410236.2016.1168000

Vick, L. R., Soporan, T., Lewis, D. R., & Zurn, J. B. (2012). Hybrid Browser/Server Collection of Streaming Social Media Data for Scalable Real-Time Analysis. *AAAI Technical*

Report, *12*(2), 29–33. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPDFInterstitial/4787/5085%5Cnpapers2://publication/uuid/68C87DD7-5174-4851-B19A-933176A5B193