iS
CHANNEL

Social Media as a Data Source in the Public Sector: Socio-Technical Challenges for Economic Indicators

Keisuke Idemitsu

MSc in Information Systems and Digital Innovation
Department of Management
London School of Economics and Political Science

KEYWORDS

Big data
Government
Public sector
Social media
Economic indicator

ABSTRACT

Social media has been used as both a communication tool and a policy evaluation tool in the public sector. Moreover, it has gradually become an important source for big data analytics in policymaking. However, little academic research has focused on the socio-technical problems that stem from the use of social media as a source for big data analytics in the public sector. This paper sheds light on such problems by focusing on several economic development indicators from social media analytics in the Japanese government. The author analyses three types of challenges identified in previous literature on big data analytics: (1) governance and privacy, (2) organisational settings, and (3) quality and bias of the data. The analysis reveals that the Ministry of Economy, Trade and Industry (METI) has tackled the problems by (1) collecting data from a study group, (2) involving academically and/or industrially highly skilled professionals across the public and the private sectors, and (3) giving explanations for the developed indicator. This paper concludes with some recommendations for other governments.

1. Background/Literature Review

Big data has been regarded as the new 'oil' in the private sector (Bhageshpur, 2019). Recently, many enterprises have started to use social media not only as a communication tool but also as a data source for big data analytics in many business fields (He et al., 2013). In the public sector, however, the potential of social media as a data source has just started to be identified. This paper aims to discover socio-technical problems that arise when the public sector uses social media as a source of big data analytics in policymaking.

Social media has been primarily used by the public sector as a communication tool among citizens. Here, researchers have focused on its function to reflect public opinion and contribute to democratic processes. They have pointed out that social media does not sufficiently represent citizens due to the digital divide and increasing amount of fake information. Nevertheless, it supports the public sector to collect opinions in real time (Desouza & Jacob, 2017). Subsequently, social media has been used as a policy evaluation tool, with which the public sector can measure the effectiveness of services quantitatively. For instance, Agostino & Arnaboldi (2017) propose a way of measuring public service effectiveness using Twitter data.

However, there is little research on how the public sector can use social media as a source for big data analytics. Desouza & Jacob (2017) point out the potential of social media as a prediction tool in the policymaking process, but empirical analysis from a socio-technical perspective is still needed to reveal the challenges for such use in the public sector (Vydra & Klievink, 2019). Regarding big data analysis in general, many researchers have addressed socio-technical problems in the public sector. Three types of concerns have been identified in the past literature. First, privacy and security are critical due to the need to collaborate across multiple agencies (Desouza & Jacob, 2017; Höchtl et al., 2016; Pencheva et al., 2018). Second, organisational setup, including the lack of capabilities for data analytics, matters in the implementation (Höchtl et al., 2016; Pencheva et al., 2018). Third, data quality and bias might cause problems (Desouza & Jacob, 2017; Höchtl et al., 2016). These points might cause inappropriate and inefficient use of big data in the public sector, and whether these problems are common in social media use as a source of big data analytics in the public sector is the theme of this paper.

The author selected economic prediction in the public sector as an example of social media data analytics, because economic prediction is an influential policymaking field directly affecting national economic policy (Blazquez & Domenech, 2018). Some research has previously introduced the methodology from a technological forecasting perspective; for example, Indaco (2018) indicates that Twitter data may be used to measure country-level gross domestic

Corresponding Author
Email Address: idemitsu1101@gmail.com

product (GDP) in a more timely manner compared to conventional estimations.

In addition, according to some reports from practitioners, several governments have attempted to implement social media data analytics as a source of economic prediction. For example, the Australian government tries to extract skills and competencies data from LinkedIn to understand the dynamics of the labour market (World Bank, 2017). However, scant academic research analyses these items. Therefore, this paper addresses the socio-technical problems in sourcing social media data for big data analytics, using the example of economic prediction. The remainder of the paper consists of four sections. Section 2 introduces the research design, section 3 describes a use case, section 4 analyses the challenges, and the final section concludes this paper.

2. Research Design

The research question of this paper is: 'What are the socio-technical challenges in social media use as a source of big data analytics in the public sector?' In order to explore the question, this paper adopts a case study approach (Flick, 2014) which enables researchers to empirically investigate the details of a particular phenomenon (Yin, 2014). The author selected a case in the Japanese government, because it was one of the earliest attempts of developing new economic indicators based on social media data analysis.

The author collected data from open source documentation, such as official websites and government reports, and analysed the three socio-technical challenges raised in the literature, as explained previously (Desouza & Jacob, 2017; Höchtl et al., 2016; Pencheva et al., 2018).

3. Case Description

The Ministry of Economy, Trade and Industry (METI) in Japan has developed several economic indicators by utilising big data since 2014 (METI & PwC Aarata LLC, 2017). In 2016, METI conducted a project "to complement, expand, and refine existing government statistics, and to develop indicators that are more prompt and accurate than existing statistics" (METI & PwC Aarata LLC, 2017, p. 7). In particular, METI commissioned a private securities company to develop an index of economic situation as a demonstration project. At the same time, an expert study group was formed to accompany the development and utilisation of the indicators.

As a result of the project, METI and the agency developed a model to estimate the Index of Industrial Production (IIP) using metrics from Twitter and blogs. According to the report (METI & PwC Aarata LLC, 2017), they first selected about 200 keywords that were considered to be strongly related to the macro index. For instance, the word "overwork" was considered to be a keyword because it represents the increase of production in manufacturing. Then, they measured the correlation between the frequency of these words and the statistical index, such as IIP. Next, they analysed the text in the documents that contained those words and included only meaningful tweets (e.g. "I overworked today."), excluding unrelated ones (e.g. "I didn't overwork today."). The obtained values were used to create a time series model in order to compare with IIP (METI & PwC Aarata LLC, 2017, p. 40).

This analysis included two essential methods: text mining and sentiment analysis. Text mining is "the systematic analysis of large-scale text collections" (Grimmer & Stewart, 2013, p. 268). It usually employs the bag-of-words model (Zhang et al., 2010) to categorise words, pre-process the text and create a document-feature matrix, which represents the frequency of each word in the whole text. The second method, sentiment analysis, estimates people's sentiment from the text (Pang & Lee, 2008). Sentiment analysis is usually conducted by

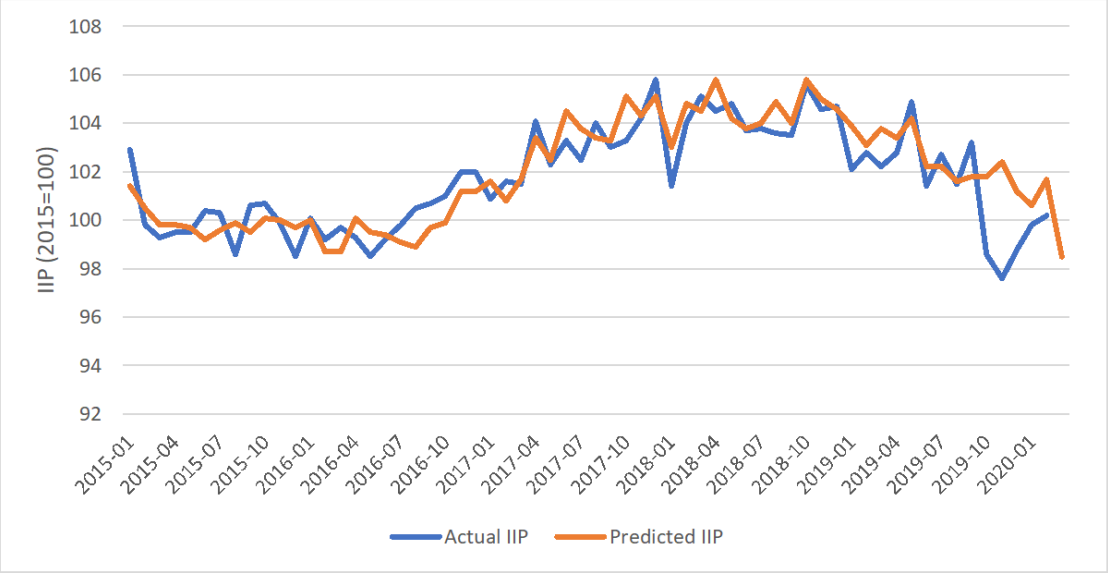

Figure 1. Comparison between actual IIP and predicted IIP (data source: METI (2020); Nomura Security Co., Ltd. (2020))

extracting the frequency of sentiment keywords (i.e. using a dictionary of sentiment keywords with the document-feature matrix made in the text mining). By combining these two methods, METI predicted real-time economic situations (see Figure 1).

METI concluded that metrics using Twitter and blogs were able to estimate the IIP with higher accuracy than those that did not use them. They emphasised the advantages, i.e. this estimation could be updated on a daily basis and could be used as a quick policy decision and investment indicator. Also, the accuracy is significantly higher than existing indices, which could improve the accuracy and speed of IIP predictions. On the other hand, METI pointed out that the model using only Twitter data was not accurate than that using multiple sources including blog data (METI & PwC Aarata LLC, 2017, p. 51).

4. Analysis

This paper analyses the case concerning the three challenges known from literature as explained in section 1: (1) governance and privacy, (2) organisational settings, and (3) data quality and bias.

First, the privacy issue is critical in social media data because the data is rarely anonymised. Further, the authors do not assume that their microblogs are mined by the public sector, although those microblogs are open to the public as long as the author does not opt out (Benedikt & Tew, 2019). Social media companies explicitly define usage rights for public entities in order to avoid inappropriate use. For instance, the Twitter Developer Agreement Policy prohibits the sharing of content with "Government End Users, whose primary function or mission includes conducting surveillance or gathering intelligence" (Twitter.com, 2020).

METI cleared this point by concluding a contract with the data collection companies. The companies formatted the raw data and passed it on to METI, which required those companies to follow the privacy rules in the contract. However, METI had to balance the conflicting goals of securing flexibility of data analysis, e.g. direct access for data source and data privacy. The company 'owning' the social media data conducted an analysis of social media indicators in the project, and denied raw data access due to security reasons (METI & PwC Aarata LLC, 2017; p. 233). This inflexibility is one of the limitations of social media analytics for public entities.

Second, social media analytics, like big data analytics in general, requires expertise. Academic research points out the lack of skills and human resources within the public sector. Therefore, METI partnered with several companies in the project. In its project report, METI indicates that the objective of the project was not only for the government to help making prompt and accurate economic and policy decisions, but also for the private sector to make quick and appropriate management decisions (METI & PwC Aarata LLC, 2017, p. 7). This second objective enabled METI to involve private sector companies in order to gain access to the expertise needed for social media analysis.

However, the project contract was limited to one year due to budget constraints. This meant that METI needed to renew the contract annually. Occasionally, an indicator is completely dependent on a company's technology. Thus, transferring the technology to the government may create a problem of intellectual property rights given that the indicator will be formally published in the following year or later. For instance, one METI indicator was dependent on a technique to extract target users from Twitter, which could only be performed by one special company. While the indicator was not used in the end, this problem would have occurred, if it had been applied as an official indicator.

Finally, quality and bias of data are problematic in social media analytics as well as in big data analytics in general. In terms of quality, METI mentioned in the report that "the model using only social media cannot be estimated with high accuracy; thus, we will consider combining multiple models to improve accuracy" (METI & PwC Aarata LLC, 2017, p. 51). Also, METI mentioned that there was missing data due to the error of social media data collection (ibid, p. 51). Further, as real-time tweets were collected via application programming interfaces (APIs) from Twitter, some data could not be verified later. The public sector needs to overcome this uncertainty by enhancing resilience of data collection, if they use social media data as an official statistic.

Moreover, a precise explanation is needed when the government releases the outcome of social media analytics. Although social media analysis can provide strong evidence of economic dynamism, governments tend to fail to provide in-depth reasons for policy choices, as some researchers pointed out in the research of evidence-based policymaking (De Marchi et al., 2016). Some biases are inevitable even if the developed indicator seems to represent the trend of markets, because social media users are not representative of the full population (Desouza & Jacob, 2017). For instance, there is a veracity issue in geolocation data on Twitter, since only ca. 20% of all tweets are tagged by geolocation data (Benedikt & Tew, 2019). Therefore, governments need to be cautious about reliability when they use indicators derived from social media data analysis.

5. Conclusion

This paper addressed the research gap and socio-technical problems of social media data mining and analysing in the public sector. It analysed a case of METI to show risks and opportunities. There are three lessons from METI's case for practitioners.

Firstly, governments need to consider privacy issues when they collect social media data. One possible solution is making a privacy contract with agency companies, as METI did in their project. METI evaluated alternatives of collecting data by establishing a study group. As such, governments will be required to consider possible choices and select the most appropriate way of data collection when they use social media data as a data source of big data analytics.

Secondly, sharing a common goal with partner companies and academic institutions is critical to complement necessary skills and knowledge for social media analysis. In general, the public sector has budgetary constraints on hiring highly skilled professionals, but METI became a platform to develop economic indicators by raising the common goal across the private sector. The example indicates the opportunities for governments to become a platform of big data analytics as well as social media analytics.

Finally, a clear explanation is necessary to build trust with citizens and businesses when the government delivers the project to the public. The developed indicators are open to the public in METI's official website with explanations about the accuracy and potential biases. Such explanations might be needed, if other governments release the results of social media analytics.

This paper concludes with limitations and the future research direction. First, the project was conducted in 2016 and 2017. Thus, further research is needed based on the recent development of technologies. Second, METI's project was outsourced to several companies. Hence, there might be different socio-technical problems in other forms of project management. Finally, the main source of social media data was Twitter. However, different social media might have different characteristics, which could cause distinct kind of socio-technical problems. Further research is needed from these points of view in the future.

References

Agostino, D., & Arnaboldi, M. (2017). Social media data used in the measurement of public services effectiveness: Empirical evidence from Twitter in higher education institutions. Public Policy and Administration, 32(4), 296–322. https://doi.org/10.1177/0952076716682369

Benedikt, L., & Tew, E. (2019, February 5). Can social media data improve official statistics? Not yet, suggests new work on tourism | National Statistical. https://blog.ons.gov.uk/2019/02/05/can-social-media-data-improve-official-statistics-not-yet-suggests-new-work-on-tourism/

Bhageshpur, K. (2019, November 15). Council Post: Data Is The New Oil -- And That's A Good Thing. Forbes. https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. Technological Forecasting and Social Change, 130, 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

De Marchi, G., Lucertini, G., & Tsoukiàs, A. (2016). From evidence-based policy making to policy analytics. Annals of Operations Research, 236(1), 15–38. https://doi.org/10.1007/s10479-014-1578-6

Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. Administration & Society, 49(7), 1043–1064. https://doi.org/10.1177/0095399714555751

Flick, U. (2014). An introduction to qualitative research (5th ed.). Sage Publications.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. Journal of Organizational Computing and Electronic Commerce, 26(1–2), 147–169. https://doi.org/10.1080/10919392.2015.1125187

Indaco, A. (2018, July 12). From Twitter to GDP: Estimating Economic Activity From Social Media. Proceedings of the 2nd International Conference on Advanced Research Methods and Analytics (CARMA 2018). CARMA 2018 - 2nd International Conference on Advanced Research Methods and Analytics. https://doi.org/10.4995/CARMA2018.2018.8316

METI. (2020, March 31). Index of Industrial Production (production / shipment / inventory, production capacity / occupancy rate), manufacturing industry production forecast index (鉱工業指数(生産・出荷・在庫、生産能力・稼働率)、製造工業生産予測指数). https://www.meti.go.jp/statistics/tyo/iip/result-1.html

METI, & PwC Aarata LLC. (2017). 平成28年度 IoT推進のための新産業モデル創出基盤整備事業 （ビッグデータを活用した新指標開発事業） 報告書 (FY2016 New Industrial Model Creation

Infrastructure Development Project for Promoting IoT (New Index Development Project Utilizing Big Data) Report).

Nomura Security Co., Ltd. (2020, March 31). SNS×AI IIP prediction index (SNS×AI 鉱工業生産予測指数). http://qr.nomura.co.jp/quants/sns_ai/#jump4

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. now Publishers Inc.

Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2018). Big Data and AI – A transformational shift for government: So, what next for research? Public Policy and Administration, 095207671878053. https://doi.org/10.1177/0952076718780537

Twitter.com. (2020, March 10). Developer Agreement and Policy – Twitter Developers. https://developer.twitter.com/en/developer-terms/agreement-and-policy

Vydra, S., & Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate. Government Information Quarterly, 36(4), 101383. https://doi.org/10.1016/j.giq.2019.05.010

World Bank. (2017). Big Data in Action for Government: Big Data Innovation in Public Services, Policy and Engagement (English). World Bank Group. http://documents.worldbank.org/curated/en/176511491287380986/Big-data-in-action-for-government-big-data-innovation-in-public-services-policy-and-engagement

Yin, R. K. (2014). Case study research: Design and methods (Fifth edition). SAGE.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 1(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0