Journalists' Use of Social Media to Infer Public Opinion: The citizens' perspective

Elizabeth Dubois, University of Ottawa
Anatoliy Gruzd, Ryerson University
Jenna Jacobson, Ryerson University

Journalists increasingly use social media data to infer and report public opinion by quoting social media posts, identifying trending topics, and reporting general sentiment. In contrast to traditional approaches of inferring public opinion, citizens are often unaware of how their publicly available social media data is being used and how public opinion is constructed using social media analytics. In this exploratory study based on a census-weighted online survey of Canadian adults (N=1,500), we examine citizens' perceptions of journalistic use of social media data. We demonstrate that: (1) people find it more appropriate for journalists to use aggregate social media data rather than personally identifiable data; (2) people who use more social media are more likely to positively perceive journalistic use of social media data to infer public opinion; and (3) the frequency of political posting is positively related to acceptance of this emerging journalistic practice, which suggests some citizens want to be heard publicly on social media while others do not. We provide recommendations for journalists on the ethical use of social media data and social media platforms on opt-in functionality.

Keywords: public opinion, social media, journalism, data journalism, ethics

Introduction

Journalists increasingly use social media data to infer public opinion, which in turn guides future reporting, prioritizes policy development, and helps citizens[i] conceptualize and feel part of their community. The reporting of public opinion may also influence citizens' opinions, which consequently shifts public opinion in a recursive loop.

Advances in social media analytics make citizens' perceptions of journalists' use of social media data particularly important. Social media analytics frequently focuses on specific events and social media platforms, yet the reporting of public opinion is often presented as representative of a wider public (Jungherr, Schoen, Posegga, & Jürgens, 2017). The reliance on social media data to infer public opinion affords an expeditious understanding of the "public's" opinions; however, the opinions are highly contextual and limited to specific publics—such as those on Twitter or those watching a political debate (Anstead & O'Loughlin, 2015). Citizens are left to assume that a given report applies broadly, or they must imagine an overarching public opinion by reflecting on various reports alongside their personal experiences (Anstead & O'Loughlin, 2015). Consequently, citizens' assessments of social media data use for inferring public opinion is crucial to understanding the value and implications of these digital journalistic practices.

There are privacy considerations that may influence citizens' perceptions of this practice. Unlike traditional polling, citizens on social media may not know when they are contributing to public opinion (Anstead & O'Loughlin, 2015). Citizens' social media posts can be used out of context and individuals can be publicly identified without their knowledge or consent, which could lead to feelings of privacy invasion, discomfort, and diminished trust in those employing these practices. Laufer and Wolfe (1977) argued that the concept of privacy needs to be tied to concrete situations that are experienced in everyday life. Shifted to social media, a case-by-case assessment is needed to determine whether the use of social media data is “reasonable, fair and ethical” (Kennedy, Elgesem, & Miguel, 2017, p. 5). Some scholars advocate for a social contract approach to develop, acknowledge, and protect privacy norms that are situated within specific contexts (Martin, 2016). Notably, Nissenbaum's (2011) concept of contextual integrity contends that individuals do not have a true choice in making privacy decisions; as such, there is a need to establish “context-specific substantive norms” that identify the conditions of acceptable social media use with a strong ethical grounding (p. 32).

Scholars have argued that ethics need to be deeply considered even when working with public data (boyd & Crawford, 2012) and that concerns over privacy are more critical when using data from online social networks (Zimmer, 2010). Institutional Review Boards often review academic proposals using social media data to ensure a high standard of ethics (Moreno, Goniu, Moreno, & Diekema, 2013); in contrast, those outside the academy, such as journalists, largely do not have the same regulations in place. Further, there is a journalistic responsibility to fairly represent the public and to provide balanced reporting, which Anstead and O’Loughlin (2012) argue must extend to social media data use for inferring public opinion. The ethics must be considered as new practices in digital journalism emerge.

Our exploratory study responds to Cohen’s (2018) call to analyze the “normalization of journalists’ reliance on social media platforms” (p. 7). 94% of online Canadians have at least one social media account (Gruzd, Jacobson, Mai, & Dubois, 2018) and journalists continue to leverage new digital techniques to assess public opinion; yet, there is little empirical research that shows how citizens feel about their social media data being used to infer public opinion. This study aims to fill this gap by analyzing whether Canadians think it is (1) appropriate, (2) useful, and (3) possible for journalists to infer public opinion from social media data.

We examine the differences in citizens’ perceptions of journalistic use of social media data for inferring public opinion based on data type and social media use. Using a survey of 1,500 online Canadians, we find journalists quoting social media posts is perceived as less appropriate than reporting aggregate data, such as trending topics or sentiment. Increased acceptance of these journalistic practices is linked to having more social media accounts and political postings, which suggests some citizens want to be represented in journalistic accounts of public opinion, while others may not. We recommend that journalists avoid simply quoting an individual’s social media posts. Journalists should be explicit in describing their data analysis procedures by identifying what, when, and how the data was collected and analyzed. The research proposes and validates a new measurement scale to assess the appropriateness of journalistic use of social media data.

The democratic value of public opinion

In 1936, Gallup introduced polling procedures that used a representative sample of the American population to predict the presidential election (Berelson, 1952). While the definition of public opinion is widely understood as an aggregate of individual opinions (Lippmann, 1922), some believe public opinion should be viewed as the outcome of individuals conversing and deliberating, which has led to the development of deliberative polling (Fishkin, 1995). Others contend public opinion can also be conceived as a form of control (Scheufle & Moy, 2000). Others suggest there are many publics and, in the majority of cases when public opinion is reported, it is the opinion of a specific public (Anstead & O'Loughlin, 2015; Gillespie, 2014).

To evaluate strategies for inferring public opinion, we consider the democratic context. The democratic role of public opinion is twofold. First, public opinion is both an outcome of and input to an informed citizenry—something which is a base requirement for democracy, according to most theories of democracy (Berelson, 1952; Dahl, 2000). Citizens use public opinion combined with other information sources and their own experiences to develop opinions about political issues and decide whether to share their opinions. For example, individuals make decisions about what opinions to share based on what they perceive the majority general public opinion to be (Noelle-Neumann, 1993; Hampton et al., 2014).

Second, public opinion provides a way citizens relate to their political system and political elites. By examining public opinion, the government and politicians can develop policy and determine strategies (Stieglitz & Dang-Xuan, 2013) that respond to the views of citizens (Erikson & Tedin, 2015), and citizens can, correspondingly, see themselves represented in and by their government's actions. This use of public opinion is not unidirectional; political elites can influence the public's political opinions to generate support for their initiatives (Leeper & Slothuus, 2014; Savigny, 2002).

Since the public predominantly sees reports of public opinion in journalistic content, the use of social media data in journalism is of particular concern. Journalists can use public opinion to provide a mirror to citizens, support and critique political elites, and defend the legitimacy of their reporting (Silverstone, 2007).

Traces of public opinion

There are multiple ways journalists depict public opinion beyond polls. Anstead and O'Loughlin (2015) identify three key ways social media data is used in journalism to report public opinion. We adopt their typology to examine citizens' perceptions of appropriateness of journalistic use of social media data to infer public opinion based on the type of social media data.

First, journalists present the citizen's voice to offer an opportunity for individuals to articulate their views in their own words (e.g., vox pop interviews). On social media, this is akin to directly quoting an individual's posts (Anstead & O'Loughlin, 2015; Broersma & Graham, 2012). We operationalize this data type as *Quote*.

Second, journalists report public reactions and responses at a general level (e.g., reporting voter turnout). On social media, this is akin to reporting the number of posts or trending topics (Anstead & O'Loughlin, 2015). We operationalize this data type as *Trend*.

Third, journalists use straw and opinion polling to explain and measure different opinions among the public (e.g., the feeling towards a political leader as positive or negative). On social media, "semantic polling" involves collecting large amounts of social media data and quantitatively reporting results, such as tone of public opinion using numbers and graphs (Anstead & O'Loughlin, 2012; 2015). We operationalize this data type as *Sentiment*.

Citizens' perceptions of journalistic use of traces of public opinion

Citizens' perceptions of journalists' work and the journalism industry have an impact on journalists' conduct (Lowrey & Anderson, 2005). As journalistic practices evolve so may perceptions of journalists' work, which can consequently impact content production.

Concurrently, how an individual perceives their social media reality—the understanding of their audience and how their audience, social media platforms, and third parties can use their data—has implications for what citizens believe is appropriate social media behaviour and appropriate social media data use (Marwick & boyd, 2011; Vitak, Blasiola, Patil, & Litt, 2015). Consequently, citizens' perceptions of journalistic use of social media data are important.

With traditional representations of public opinion, respondents are typically aware of why they are asked their opinion and by whom. Conversely, social media posts are largely not intended for journalistic use and consent is likely not requested. This distinction is particularly relevant for the journalistic practice of quoting social media posts as this practice often identifies the individual being quoted (e.g., publishing the username), which could be perceived as an invasion of privacy.

As producers of social media data, individuals have a unique perspective as to what their data means and how it should be interpreted. Journalists, however, can interpret the data without this perspective, which could lead to a misrepresentation of the public's wants and needs. Furthermore, people may adjust their practices when they witness social media data being used (boyd, 2014; Couldry, Fotopoulou, & Dickens, 2016). Perceptions of these journalistic practices could impact what version of public opinion journalists are able to capture.

To complicate matters, the utility of using social media to gauge public opinion is debated (Jungherr et al., 2017) and many contend this form of public opinion is limited, contextual, and multiple (Anstead & O'Loughlin, 2015; Gillespie, 2014; Jungherr et al., 2017). Citizens may question the value and integrity of the journalists' reports if they are unsure whether journalists can use social media to reliably make claims. Considering the democratic value of public opinion, citizens need to be able to trust—or at least, know how to evaluate—the information they use to develop their own opinions. Media representations of public opinion also need to be adequately contextualized so individuals can see themselves represented in their political system and feel confident that political elites are making decisions based on trustworthy information.

In addition to the three measures of appropriateness by data type (*Quote*, *Trend*, and *Sentiment*), we investigate whether citizens believe journalists *Can* infer public opinion from social media data, whether they *Should*, and whether it is *Useful*. By understanding the extent to which these slightly different frames are meaningfully different to our respondents, we provide a nuanced understanding of the public's perceptions towards the journalistic use of social media data and develop a measurement scale that can be used in future research.

Privacy concerns of personally identifiable social media posts

While there is limited research on the perceptions of journalistic use of social media data to infer public opinion, there is extensive research on privacy concerns related to social media data. Privacy is contextual (Nissenbaum, 2011) and individuals have different perceptions of privacy when data is used out of context or by third parties. The third-party use of personally identifiable information on social media can be problematic even if the data is publicly available (Gruzd, Jacobson, & Dubois, 2017). There are technical and ethical challenges as re-identification of anonymized data is often possible (Zimmer, 2010). For example, when a social media post is directly quoted with the username removed, others can simply search the text online and re-identify the author.

Some social media data types may elicit more concern than others. Writing before the widespread use of social media, Ackerman, Cranor, and Reagle (1999) proclaimed that "not all data is the same" (p. 3), as people's comfort with sharing information depends on the data type. Shifted to social media, we hypothesize people will express varying levels of concern and comfort depending on the specific data type:

H1: Individuals will consider it is more appropriate for journalists to use aggregated social media data (Trend and Sentiment) as compared to personally identifiable social media data (Quote).

Social media use and political opinion sharing

Social media use is likely a factor that influences an individual's perception of journalistic use of social media data; however, the direction of the relationship is unclear. Extending the privacy discussion, increased social media use could lead to lower acceptance of social media data use by journalists because the individual has more to lose. Alternatively, increased social media use could lead to higher acceptance of this journalistic practice for at least two reasons. First, increased use of digital tools increases media literacy (Livingstone, 2004); accordingly, individuals who are cognizant of how their data could be used may modify their behaviour to limit their risks (Couldry & Powell, 2014), which may increase their acceptance of the practice. Second, some individuals may specifically use social media as a platform to have their opinions heard and/or to influence political agendas and public opinion (Bennett, 2012; Dubois, 2015). Considering the ambiguity in the literature on the impact of general social media use, we pose the following research questions:

Does an individual's social media use in terms of the number of social media accounts (**RQ1a**) and frequency of posting on social media (**RQ1b**) impact their perception of journalistic use of social media data to report public opinion?

Previous research suggests that general social media use can lead to increased political expression online and increased political participation online and offline (Gil de Zúñiga, Molyneux, & Zheng, 2014). 33% of Canadians report having political discussions on social media (Hilderman & Anderson, 2017), and a larger percentage (48%) of Canadians use social media to gather news (Brin, 2017), which is noteworthy as news consumption has been positively linked to political expression online (Shah, Cho, Eveland, & Kwak, 2005).

It is important to specifically investigate political uses of social media because the relatively small population of users who post political content may perceive journalists' use of social media data differently than the general public. For activist purposes, people are motivated to use social media for logistics, community building, and sharing political information (Bennett & Segerberg, 2012; Chadwick & Howard, 2010; Dubois, 2015; White et al., 2015). An individual who uses social media for social and political change may be more accepting of their social media data being used by journalists. Indeed, some individuals may specifically aim to influence the issues journalists report on and the framing of these issues (Bennett & Segerberg, 2012; Chadwick, 2011; Jungherr, 2014; Papacharissi & de Fatima Oliveira, 2012).

Similarly, Dubois (2015) found that opinion leaders sourced from the Canadian political Twittersphere make decisions about the content and timing of their posts based, in part, on their audience and whether they are likely to have political influence. Influencing journalists and politicians was a motivation for posting political opinions publicly on social media; and, being retweeted, liked, or quoted by a journalist was a measure of success, which encouraged continued political postings online. As such, we hypothesize:

H2: The more an individual shares political opinions on social media, the more likely they are to positively perceive journalistic use of social media data to report public opinion.

H3: If an individual or their family or friends have been quoted in the news, the more likely they are to positively perceive journalistic use of social media data to report public opinion.

In contrast, some people intentionally choose to self-censor by not posting political content; common reasons for self-censorship include a lack of interest and a fear of upsetting friends or attracting negative responses (Dubois, 2015). The risk of a journalist using an individual's social media data to infer public opinion may be an additional factor. As such, we hypothesize:

H4: The more an individual self-censors (i.e., chooses not to post political content), the less likely they are to positively perceive journalistic use of social media data to report public opinion.

Methods

Data collection

Research Ethics approval was obtained from two Canadian universities. We collected data using a market research data company, Research Now, for panel recruitment of online Canadian adults (at least 18 years old). Hosted by Qualtrics and available in English and French, the survey was open from June 1 to July 15, 2017. After data cleaning, a total of 1,500 completed responses are included.

Quota sampling was used to increase the representativeness of the data. Participants were pre-screened based on age, gender[ii], and location[iii] to match the distributions in the 2016 Statistics Canada Census. The survey was designed and piloted over one year to ensure readability, accessibility, and a high-quality research design. Aligned with Research Now's typical recruitment process, and as a token of appreciation and compensation for participants' time, participants were given eRewards for completing the survey and the points earned could be transferred to loyalty rewards.

Variables and measurement

We included seven control variables: age, gender, education, income, employment, self-reported internet skills, and general comfort with social media use by third parties. Table 1 summarizes the variables, coding, and frequencies.

Table 1. Variables

	Coding	**Mean/N**	**S.D./%**
Independent Variables			
Age	Scale (range: 18–91)	47.94	16.261
Female	1 = female	757	50.50%
Male	1 = male	727	48.50%
Trans*, non-binary, two-spirit, genderqueer, other	1 = non-binary	12	0.80%
Education	Scale (8-point: some school, no degree–	4.22	1.439

	doctorate degree)		
Income	Scale (7-point: less than $20,000–more than $120,000)	4.27	1.886
Employment	1 = employed	1016	67.70%
Internet skill	0 = poor or fair; 1 = good or excellent	0.8987	0.30187
Comfort	Scale (nine 7-point items: extremely comfortable–extremely uncomfortable)	4.90	1.69667
# Social Media	Scale (range: 0–10)	3.64	2.322
Post frequency	Scale (6-point: never–several times a day)	1.99	1.38
Political post frequency		0.87	1.235
Quoted in news	1 = has been quoted	128	8.50%
Self-censor	1 = has self-censored	393	26.20%
Dependent Variables			
Quote	Scale (7-point: strongly disagree–strongly agree)	3.03	1.616
Trend		3.50	1.525
Sentiment		3.52	1.523
Can		3.22	1.597
Should		3.11	1.589
Useful		2.87	1.607

Our independent variables are: number of social media accounts, frequency of posting, frequency of posting political content, self-censorship, and past experience being quoted by a journalist. The number of social media accounts was constructed by counting the number of social media platforms an individual reported having[iv].

Additionally, the Comfort variable was constructed to control for general concerns about social media data use by organizations versus concerns specifically related to the journalistic use of social media data. The scale was previously developed and evaluated by Gruzd, Jacobson, and Dubois (2017). Participants were asked to score their comfort level on a 7-point Likert scale—from 1–7; 1-“extremely comfortable” to 7-“ extremely uncomfortable”—with third parties accessing their publicly available information from social media based on nine data types. The nine items were used to form a composite comfort level variable based on the mean. Before merging the items, we confirmed that all items load on a single dimension based on an exploratory factor analysis. Cronbach's Alpha of 0.97 confirmed the reliability of this comfort scale.

Six dependent variables were measured on a 7-point Likert scale that asked, “To what extent do you agree with the following statements” —ranging from 1-“Strongly disagree” to 7-“Strongly agree.” Each variable is coded so that *higher values* mean the respondent agrees more with the statement, or in other words, they perceive this data use positively. The six items are:

Can: "I think journalists can estimate how most people feel about issues by analyzing social media posts"

Should: "I think journalists should collect social media posts in order to understand how the public are responding to issues"

Useful: "I find it useful when a journalist uses social media posts in their news article or broadcast"

Quote: "I think it is appropriate for a journalist to quote tweets or other types of social media posts in their news article or broadcast"

Trend: "I think it is appropriate for a journalist to report the trending topics or overall number of posts on social media related to given issues"

Sentiment: "I think it is appropriate for a journalist to report how positively or negatively people on social media are responding to given issues on average"

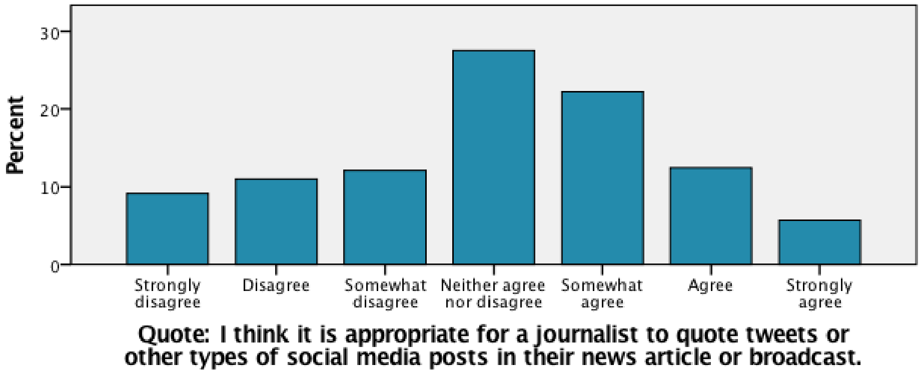
Since these dependent variables are potentially similar in what they represent, we conducted a factor analysis and found they all load to a single factor. We also performed a hierarchical regression analysis on this new variable for comparison. We used paired sample t-tests to respond to H1 and OLS hierarchical regressions to respond to the remaining research questions and hypotheses. We used SPSS for this analysis.

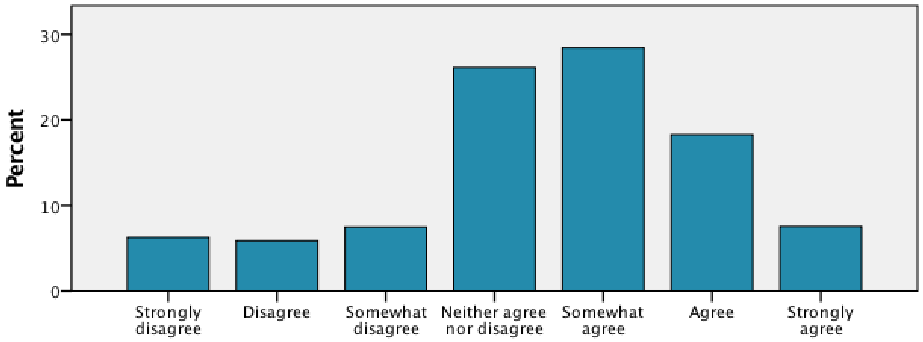
Results

Citizens' perceptions of social media data use by journalists

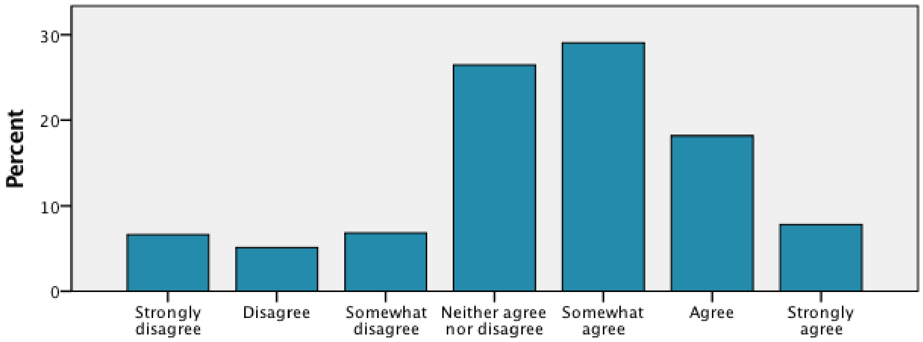
Figure 1 shows the distributions of the six dependent variables. Responses are well spread across all values and the distributions are relatively symmetric. The proportion of individuals who perceive journalistic use of social media data to infer public opinion positively versus negatively is roughly the same across all six variables with slightly more respondents agreeing (which indicates a positive view) than disagreeing.

Figure 1. Distribution of dependent variables (N=1,498)

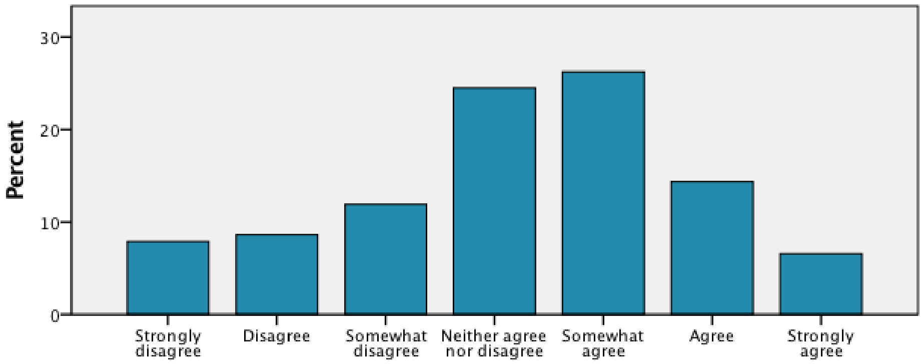




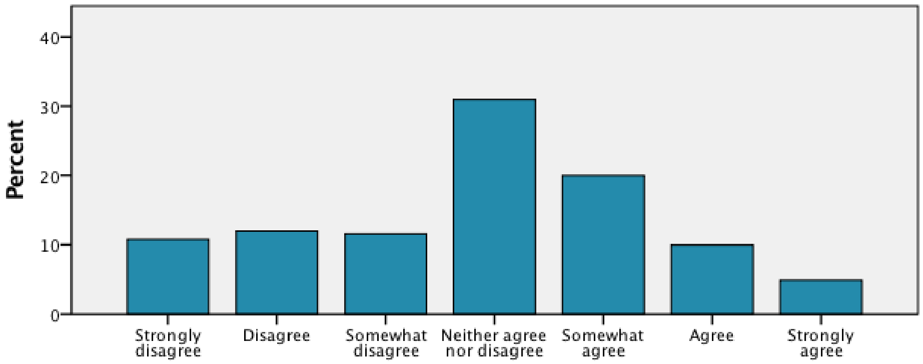
Trend: I think it is appropriate for a journalist to report the trending topics or overall number of posts on social media related to given issues.



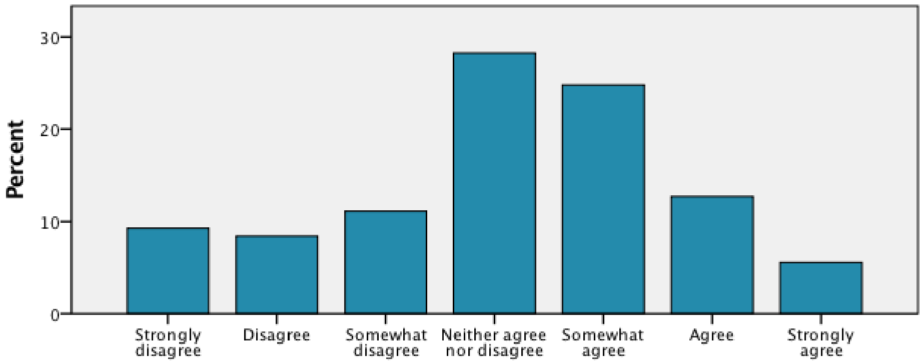
Sentiment: I think it is appropriate for a journalist to report how positively or negatively people on social media are responding to given issues on average.



Can: I think journalists can estimate how most people feel about issues by analyzing social media posts.



Useful: I find it useful when a journalist uses social media posts in their news article or broadcast.



Should: I think journalists should collect social media posts in order to understand how the public are responding to issues.

Appropriateness of data type

Our first hypothesis is that individuals will consider it is more appropriate for journalists to use aggregated social media data (Trend and Sentiment) as compared to personally identifiable social media posts (Quote). We found a significant difference in the scores for *Quote* (M=3.03, SD=1.616) and *Trend* (M=3.50, SD=1.525), t(1497)=13.321, p < 0.001 as well as in the scores for *Quote* (M=3.03, SD 1.616) and *Sentiment* (M=3.52, SD=1.523), t(1497)=14.008, p < 0.001. However, there was no significant difference between *Trend* and *Sentiment*, t(1497)=-0.804, p=0.422. Aligned with **H1**, people perceive journalists quoting social media posts as less appropriate than reporting aggregate data.

The impact of an individual's social media behaviours

To understand the impact of social media use on individuals' perceptions of journalistic use of social media data, we consider the number of social media accounts a person has, frequency of social media posting, and frequency of posting political information on social media. A correlation table is omitted given the Pearson Correlation Coefficients ranged from 0.000 to +/- 0.48 representing primarily small correlations and none that pose a substantial problem for our analysis. Tests for multicollinearity were acceptable with VIF ranges from 1 to 1.6, which is within an acceptable range (below 5).

The standardized coefficients for seven regressions with control variables only are reported in Table 2. Age is consistently significant and negative as is Comfort. Respondents who are older and respondents who are generally uncomfortable with social media data use are less likely to perceive journalistic use of social media data to infer public opinion positively. Gender is significant in two (*Trend* and *Can*) of the seven regressions, and education is significant in only one regression (*Can*). The remaining variables are not significant. All of our models are statistically significant and the adjusted R^2 values range from 6.4% to 10%, which indicates a small but significant effect size (Cohen, 1992). It is interesting to examine the change in adjusted R^2 values and significance of that change because the research analyzes the relationship between specific and theoretically-derived independent variables related to social media use and our dependent variables; further, we are not attempting to explain all variance within those dependent variables (Wampold & Freund, 1987).

Table 2. OLS Regression with control variables only

	Quote	Trend	Sentiment	Can	Should	Useful	Factor
Age	-.079**	-.065*	-.057*	-.088**	-.066*	-.077**	-.086**
Gender							
Female	-.008	.051*	.049	.064*	-.007	.043	.038
Non-binary	.034	.005	.017	.075**	.025	.036	.037
Education level	.041	.036	.018	-.080**	.029	-.002	.010
Income	.024	.008	-.015	-.004	-.036	-.017	-.008
Employed	-.006	-.016	-.008	.010	.012	-.009	-.004
Skilled	.000	.017	.011	-.003	-.013	-.012	.000
Comfort	-.250***	-.243***	-.266***	-.208***	-.264***	-.274***	-.301***
N	1487	1487	1487	1487	1487	1487	1487
R^2	0.077	0.069	0.080	0.074	0.081	0.088	0.106
Adjusted R^2	0.072	0.064	0.075	0.069	0.076	0.083	0.101

Notes: * p<0.05, ** p<0.01, *** p<0.001; OLS regressions presenting standardized beta coefficients; Omitted categories are Male, Unemployed, Unskilled.

In Table 3, we add social media use variables and the adjusted R^2 values increase between approximately 1.8 and 4.6 percentage points—ranging from 9.2% to 14.5%. Though the increases are not large, the change is significant in all models. This means our new independent variables together do significantly account for an amount of variance in each of our dependent variables above and beyond the demographic variables alone (Cohen, Cohen, West, & Aiken, 2013; Wampold & Freund, 1987). This helps establish an understanding of the role of the specific independent variables of interest. We are not concerned with explaining all variance in our dependent variable in this study; as such, it is acceptable that our adjusted R^2 values remain indicative of a small effect size (Cohen, 1992)[v].

Furthermore, there are meaningful changes in the coefficients, which highlights the value of this analysis. Age is no longer significant in any regression and gender is no longer significant in *Trend*. Both Female and Non-binary were significant in one regression (*Can*); however, when adding the social media variables, Female is no longer significant. Only a few people identified as non-binary (N=12), which makes the significance questionable. Education remains significant in one regression (*Can*).

Table 3. OLS regression with all independent variables

	Quote	Trend	Sentiment	Can	Should	Useful	Factor
Age	-.007	0.011	.003	-.011	.004	-0.001	.000
Gender							
Female	-.013	.040	.038	.048	-.012	.036	.027
Non-binary	.021	-.006	.007	.063*	.014	.020	.023
Education level	.017	.017	.007	-.098***	.005	-.027	-.014
Income	.030	.008	-.016	-.003	-.030	-.010	-.005
Employed	-.013	-.025	-.018	-.003	.005	-.020	-.015
Skilled	-.016	.000	-.004	-.023	-.028	-.028	-.019
Comfort	-.211***	-.206***	-.230***	-.165***	-.229***	-.228***	-.245***
# social media	.135***	.149***	.102***	.143***	.140***	.134***	.160***
Post frequency	-.026	.014	.054	.057	-.038	-.006	.011
Political posting	.123***	.064*	.026	.055	.118***	.108***	.098***
Quoted in news	.061*	.043	.052*	.024	.052*	.097***	.066**
Self-censor	-.020	-.007	.025	-.050*	-.030	-.050*	-.026
N	1487	1487	1487	1487	1487	1487	1487
R^2	0.116	0.100	0.101	0.110	0.117	0.136	0.153
Adjusted R^2	0.108	0.092	0.093	0.102	0.110	0.129	0.145
Change in R^2	0.039***	0.031***	0.022***	0.036***	0.036***	0.048***	0.049***

Notes: * p<0.05, ** p<0.01, *** p<0.001; OLS regressions presenting standardized beta coefficients; Omitted categories are Male, Unemployed, Unskilled.

The number of social media accounts an individual has is a significant predictor of perceptions of journalistic use of social media. Across all dependent variables, the number of social media accounts an individual has is significant and positive. As such, the response to **RQ1a** (Does an individual's general social media use impact their perception of journalistic use of social media data to report public opinion?) is yes. Respondents who use more social media are more likely to positively perceive journalistic use of social media data to infer public opinion. These respondents perceive this journalistic practice as appropriate; they think that journalists *Can* and *Should* use social media data to infer public opinion and they think this practice is *Useful*.

Next, we asked, **RQ1b**: Does an individual's frequency of posting on social media impact their perception of journalistic use of social media data to report public opinion? According to our findings, the answer is no. Across all regressions, frequency of posting on social media was not significant.

In contrast, posting political content on social media is significant and positive in five regressions. This suggests that, in general, those who make political posts more frequently find it appropriate for journalists to *Quote* social media posts and report social media *Trends*. Interestingly, posting political content is not significant when considering reporting *Sentiment* as the dependent variable. Likewise, frequency of political posting is not significant when considering whether journalists *Can* infer public opinion from social media data, although it is significant when considering whether journalists *Should* and whether it is *Useful*. We find that **H2** (The more an individual shares political opinions on social media, the more likely they are to positively perceive journalistic use of social media data) is generally, but not always, supported.

Having been quoted in the news has mixed results. In five regressions (*Quote*, *Sentiment*, *Should*, *Useful,* and *Factor*) having been quoted by a journalist in the past is significant and positive—though relatively weak. **H3** (If an individual or their family and friends have been quoted in the news, they are more likely to positively perceive journalistic use of social media data) is generally, but not always supported.

Finally, we consider those who avoid sharing on social media because journalists may use that data and find that self-censorship is significant in two (*Can* and *Useful*) of the six regressions. As expected, the relationship is negative in both cases. **H4** (The more an individual self-censors, the less likely they are to positively perceive journalistic use of social media data) is not consistently supported by our data.

Notably, the regression using our factor has a higher Adjusted R^2 and the largest difference from step one to two (about 4.4 percentage points). Future work might investigate the similarities and differences across our dependent variables to determine whether a single factor is the best approach moving forward.

Discussion

Advances in digital journalism increasingly require journalists use social media data (Cohen, 2018); furthermore, using social media to infer public opinion is becoming common practice (Anstead & O'Loughlin, 2015). Citizens' perceptions of whether these practices are *appropriate*, *useful, can,* and *should* be done (our dependent variables) help to guide ethical digital journalism. We show a number of interesting trends in the relationships between our variables. Individuals find it more appropriate for journalists to use aggregate data, rather than personally identifiable data (H1 supported). Further, as the number of social media accounts an individual has and their frequency of posting political content on social media increases so do their positive perceptions toward journalistic use of social media data to infer public opinion (RQ1a; H2 supported); however, the frequency of posting on social media is not significant (RQ1b). Having previously been quoted by a journalist (H3, somewhat supported) and engaging in self-censorship (H4, not supported) are sometimes, but not always, significant, which suggests there could be meaningful differences among the six dependent variables.

Our exploratory models—and the change in R^2 across each hierarchical regression—are all significant, which suggests there is explanatory value in the addition of our independent variables (Cohen et al., 2013). A relatively low adjusted R^2 and modest change in adjusted R^2 values across our regressions limit the explanatory power of our models and suggests that, though our independent variables are relevant, additional variables may help explain individuals' perceptions of journalistic use of social media data. This highlights the complexity of individuals' perceptions. While we did not aim to explain all the variance in our dependent variables, future work may build from our findings.

Intending to influence

Individuals who use more social media tend to be more accepting of journalists using social media data to report public opinion. This trend may capture a segment of the population who intentionally use social media to reach a wider audience and, perhaps, influence public opinion and political elites. These individuals may want to be quoted by journalists. Indeed, having one's voice heard is a motivator for people to share their political opinions (Pingree, 2007). Activists' strategies for influencing political agendas (Bennett & Segerberg, 2012; Chadwick, 2011) rely on journalists paying attention to and using their social media data.

Risks—in terms of potential harm to the individuals posting on social media—may be minimized because those who do not accept these journalistic practices do not frequently post political content that journalists can use. This argument breaks down if journalists use content that individuals do not consider to be political to infer public opinion. For example, someone could simply tweet about visiting a national park, but a journalist could use this as evidence that there is support for parks in a political debate regarding funding cuts.

Individuals have varying responses to journalistic use of social media data depending on their number of social media accounts and the frequency of political posting. It is necessary to recognize these divides when developing ethical standards for digital journalism. Individuals are ill-equipped to protect themselves without an understanding of how social media data is used. While digital journalism initiatives tend to require journalists do more with less resources and work more quickly (Cohen, 2018; Hermida & Young, 2016), we suggest there are important ethical practices for journalists' consideration. We echo calls for greater media and data literacy as well as methodological transparency in the use of publicly available social media data (Anstead & O'Loughlin, 2012; Elueze & Quan-Haase, 2018; Kennedy & Moss, 2015).

Partial publics

In "networked publics," a sense of community can develop using social media, thus forming a public for which a generalized opinion may be inferred (boyd, 2010). Social media data, however, can only tell partial stories about specific publics, which limits generalizability (Barberá & Rivero, 2015; Jungherr, 2016). We have identified one of the ways the public may be divided into multiple publics: based on the number of social media accounts. It is likely that this variable helps to identify activist-type social media users who want to be heard. Alternatively, respondents who use multiple social media platforms may also be more likely to use more public platforms where third parties commonly re-use user data, and where such uses may have been normalized and encouraged by the platform and users (e.g., Twitter).

Gillespie (2014) argues that beyond networked publics, algorithms afford the formation of “calculated publics,” which are not established based on intentional interactions among people, but based on how an algorithm organizes and categorizes people and their behaviours. Kennedy and Moss (2015) further distinguish between *known* and *knowing* public. *Known* publics are those that we can understand based on their social media data; journalists may infer the opinion of such a public by collecting and analyzing social media data. *Knowing* publics, however, offer social media users agency. By reflecting on the existence of the public, individuals become more aware and active in the construction of that public and subsequent representation of the opinion of that public (Kennedy & Moss, 2015).

As people gain awareness of digital journalistic practices, the very nature of public opinion that journalists are able to infer and report may change. Since reported public opinion tends to influence public opinion and voters’ decision-making, the integrity of elections is a crucial concern (Anstead & O’Loughlin, 2012). Similarly, people reflect on social media analytics and potentially change their behaviour (Couldry & Powell, 2014) and we, therefore, need to understand the ways people respond and interact with this data.

The ethical use of social media data by journalists requires a consideration of how different publics might respond to and interact with the version(s) of public opinion being reported. To do this, citizens require information about how journalists collect, analyze, and present social media data. Journalists must be cognizant that some individuals on social media want to be heard; these individuals can manipulate the social media system to achieve their desired outcome of having journalists receive a particular signal from the “public” they are examining. Increased critical reflection on how social media data is collected and analyzed can assist journalists and citizens identify when this vulnerability is being exploited.

The public’s perception of journalistic data use

Considering the democratic relevance of public opinion in a digital context, it is critical to incorporate citizens’ perspective into research on the ethics of social media data use. Our study provides a rich account of how individuals respond to these emerging journalistic practices. We identified that social media needs to be understood at the data type level, rather than an amorphous understanding of “social media.” Our research shows a statistically significant difference in social media users’ attitudes in cases when they—and their data—are used as the unit of the analysis (e.g., quoting a social media post) versus when they are part of a larger dataset (e.g., the sentiment of many individuals’ posts). Journalists’ use of social media data that is personally identifiable tends to be viewed as less appropriate than aggregate data types. This finding suggests journalists should avoid using personally identifiable social media data.

In all cases, the number of social media accounts an individual has and the frequency of political posting are significant, which suggests a level of consistency across measures. While most of the regressions produced similar results, the dependent variable *Can* was divergent. Media literacy may be a factor because education was significant and strong in both the control variable model and when the social media variables were added.

Our research points to an interesting trend which suggests some citizens want to be heard on social media while others do not. Accordingly, journalists should get individuals’ consent—especially when relying on individual-level data as opposed to aggregate data. When journalists rely on a large dataset, it may be impractical for journalists to contact users for consent. As such, social media platforms should implement user-friendly functionality for users to declare whether and how their data can be used by third parties.

Notably, general comfort with third parties using social media data is one of the strongest and consistent predictors, which suggests that people who are uncomfortable with journalistic use of social media data are also generally uncomfortable with any third party using their data. For some, the concern appears to be rooted in the use of the social media data, rather than who is using that data. As such, the recommendation for social media platforms to incorporate technical changes would have utility beyond journalism and would afford an opt-in process for users' data being used by third parties. We need to move beyond a de facto presumption of acceptance merely because the data is public.

Acknowledgments

This research is supported in part through a five-year initiative on "Social Media Data Stewardship" funded by the Canada Research Chairs program (2015–2020; Principal Investigator: Gruzd, A), eCampusOntario research funding and the Ted Rogers School of Management at Ryerson University. The authors would like to thank Jordan Kilfoy, Christine Gagnon, and Jocelyn Stéphane Cadieux for their help with the French translation of the survey and members of the Social Media Lab at Ryerson University for their feedback on the survey design. The authors would also like to thank Alexander Furnas for his help and feedback during the preparation of this manuscript.

References

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce* (pp. 1–8). ACM.

Anstead, N., & O'Loughlin, B. (2012). Semantic polling: The ethics of online public opinion. *LSE Media Policy Project: Media policy brief 5.* Retrieved from: http://eprints.lse.ac.uk/46944/1/LSEMPPBrief5.pdf.

Anstead, N., & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication, 20*(2), 204–220.

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729.

Bennett, W. L. (2012). The personalization of politics: Political identity, social media, and changing patterns of participation. *The ANNALS of the American Academy of Political and Social Science*, *644*(1), 20–39.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication & Society*, *15*(5), 739–768.

Berelson, B. (1952). Democratic theory and public opinion. *Public Opinion Quarterly*, *16(*3), 313–330.

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *Networked self: Identity, community and culture on social network sites.* New York, NY: Routledge.

boyd, d. (2014). *It's complicated: The social lives of networked teens*. New Haven, CT: Yale University Press.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Brin, C. (2017). Canada. *Reuters Institute Digital News Report*. Retrieved from http://www.digitalnewsreport.org/survey/2017/canada-2017/

Broersma, M., & Graham, T. (2012). Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice*, *6*(3), 403–419.

Chadwick, A. (2011). The political information cycle in a hybrid news system: The British Prime Minister and the 'bullygate' affair. *International Journal of Press/Politics, 16*(1), 3–29.

Chadwick, A., & Howard, P. N. (Eds.). (2010). *Routledge handbook of internet politics*. Abingdon, Oxon: Taylor & Francis.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155–159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. London, UK: Routledge.

Cohen, N. S. (2018). At work in the digital newsroom. *Digital Journalism.* DOI: 10.1080/21670811.2017.1419821

Couldry, N., & Powell, A. (2014). Big data from the bottom up. *Big Data & Society*, *1*(2), 1–5.

Couldry, N., Fotopoulou, A., & Dickens, L. (2016). Real social analytics: A contribution towards a phenomenology of a digital world. *The British Journal of Sociology*, *67*(1), 118–137.

Dahl, R. A. (2000). *On democracy.* New Haven, CT: Yale University Press.

Dubois, E. (2015). The strategic opinion leader (PhD thesis). University of Oxford.

Elueze, I., & Quan-Haase, A. (2018). Privacy attitudes and concerns in the digital lives of older adults: Westin's privacy attitude typology revisited. *American Behavioral Scientist*. Retrieved from: https://arxiv.org/pdf/1801.05047.pdf

Erikson, R. S., & Tedin, K. L. (2015). *American public opinion: Its origins, content and impact*. New York, NY: Routledge.

Fishkin, J. S. (1995). *The voice of the people: Public opinion and democracy*. New Haven, CT: Yale University Press.

Gearhart, S., & Zhang, W. (2014). Gay bullying and online opinion expression: Testing spiral of silence in the social media environment. *Social science computer review*, *32*(1), 18–36.

Gil de Zúñiga, H., Garcia-Perdomo, V., & McGregor, S. C. (2015). What is second screening? Exploring motivations of second screen use and its effect on online political participation. *Journal of Communication*, *65*(5), 793–815.

Gil de Zúñiga, H., Molyneux, L., & Zheng, P. (2014). Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. *Journal of Communication*, *64*(4), 612–634.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). Cambridge, MA: MIT Press.

Gruzd, A., Jacobson, J., & Dubois, E. (2017). You're hired: Examining acceptance of social media screening of job applicants. *Proceedings of the 23rd Americas Conference on Information Systems*. Boston, MA. Available at http://aisel.aisnet.org/amcis2017/DataScience/Presentations/28/

Gruzd, A., Jacobson, J., Mai, P., & Dubois, E. (2018). The state of social media in Canada. *Social Media Lab*. Retrieved from https://doi.org/10.5683/SP/AL8Z6R

Hampton, K., Rainie, L., Lu, W., Dwyer, M., Shin, I., & Purcell, K. (2014, August). Social media and the 'spiral of silence' *Pew Research Center.* Retrieved from http://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/

Hermida, A., & Young, M. L. (2016). Finding the data unicorn: A hierarchy of hybridity in data and computational journalism. *Digital Journalism*, *5*(2), 159–176.

Hilderman, J., & Anderson, K. (2017). *Democracy 360* Samara. Retrieved from http://www.samaracanada.com/docs/default-source/Reports/samara's-2017-democracy-360.pdf?sfvrsn=16

Johnson, T. J., & Kaye, B. K. (2015). Site effects: How reliance on social media influences confidence in the government and news media. *Social Science Computer Review*, *33*(2), 127–144.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239–259.

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, *13*(1), 72–91.

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, *35*(3), 336–356.

Kennedy, H., & Moss, G. (2015). Known or knowing publics? Social media data mining and the question of public agency. *Big Data & Society*, *2*(2), 1–9.

Kennedy, H., Elgesem, D., & Miguel, C. (2017). On fairness: User perspectives on social media data mining. *Convergence: The International Journal of Research into New Media Technologies*, *23*(3), 270–288.

Laufer, R. S., & Wolfe, M. (1977). Privacy as a concept and a social issue: A multidimensional developmental theory. *The Journal of Social Issues*, *33*(3), 22–42.

Lee, S., Chung, J. E., & Park, N. (2016). Linking cultural capital with subjective well-being and social support: the role of communication networks. *Social Science Computer Review*, *34*(2), 172–196.

Leeper, T. J., & Slothuus, R. (2014). Political parties, motivated reasoning, and public opinion formation. *Political Psychology*, *35*(S1), 129–156.

Lippmann, W. (1922). *Public opinion*. New York, NY: Harcourt, Brace and Company.

Livingstone, S. (2004). Media literacy and the challenge of new information and communication technologies. *Communication Review*, 1(7), 3–14

Lowrey, W., & Anderson, W. (2005). The journalist behind the curtain: Participatory functions on the internet and their impact on perceptions of the work of journalism. *Journal of Computer-Mediated Communication*, 10(3).

Martin, K. (2016). Understanding privacy online: Development of a social contract approach to privacy. *Journal of Business Ethics*, *137*(3), 551–569.

Marwick, A. E., & boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133.

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, Behavior, and Social Networking*,*16*(9), 708–713.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.

Noelle-Neumann, E. (1993). *The spiral of silence: Public opinion, our social skin.* Chicago, IL: University of Chicago Press.

Papacharissi, Z., & de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication*, *62*(2), 266–282.

Pingree, R. J. (2007). How messages affect their senders: A more general model of message effects and implications for deliberation. *Communication Theory*, *17*(4), 439–461.

Savigny, H. (2002). Public opinion, political communication and the Internet. *Politics, 22*(1), 1–8.

Scheufle, D. A., & Moy, P. (2000). Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International Journal of Public Opinion Research, 12*(1), 3–28.

Shah, D. V., Cho, J., Eveland Jr., W. P., & Kwak, N. (2005). Information and expression in a digital age: Modeling Internet effects on civic participation. *Communication Research*, *32*(5), 531–565.

Silverstone, R. (2007). *Media and morality: On the rise of the mediapolis*. Cambridge, MA: Polity Press.

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291.

Vitak, J., Blasiola, S., Patil, S., & Litt, E. (2015). Balancing audience and privacy tensions on social network sites: Strategies of highly engaged users. *International Journal of Communication*, *9*(20), 1485–1504.

Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. *Journal of Counseling Psychology*, *34*(4), 372–382.

White, B., Castleden, H., & Gruzd, A. (2015). Talking to Twitter users: Motivations behind Twitter use on the Alberta oil sands and the Northern Gateway Pipeline. *First Monday, 20*(1), Retrieved from: http://firstmonday.org/ojs/index.php/fm/article/view/5404

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325.

Author Information

Dr. Elizabeth Dubois in an Assistant Professor in the Department of Communication, University of Ottawa, Canada. She completed her doctoral work at the Oxford Internet Institute, University of Oxford. Her research focuses on political uses of digital media and political opinion formation. Email: elizabeth.dubois@uottawa.ca

Dr. Anatoliy Gruzd is a Canada Research Chair in Social Media Data Stewardship, Associate Professor and Research Director of the Social Media Lab at Ryerson University's Ted Rogers School of Management in Toronto, Canada. Gruzd studies how social media use is changing the ways in which people and organizations communicate, connect and how these changes impact our society. Email: gruzd@ryerson.ca

Dr. Jenna Jacobson is an Assistant Professor at Ryerson University's Ted Rogers School of Retail Management in Toronto, Canada. Prior to joining the School in July 2018, she was a Postdoctoral Research Fellow at Ryerson University's Social Media Lab studying how privacy, ethics, and data use are perceived by social media users in relation to their data being mined by third parties. She is also a Chair of the International Conference on Social Media & Society. She received her PhD from the University of Toronto, Faculty of Information.
Email: jenna.jacobson@ryerson.ca

Data Availability

Please contact Dr. Anatoliy Gruzd at gruzd@ryerson.ca to obtain an anonymized copy of the data used in this study for replication purposes.

[i] We use the term "citizen" in a non-legal context to refer to a member of the public.

[ii] We recognize gender is not binary; however, the question was phrased to be aligned with Statistics Canada to recruit a representative sample for statistical analysis. Participants were informed that they could later respond to a more inclusive question.

[iii] Unfortunately, Research Now does not have panel survey participants from Yukon, Northwest Territories, and Nunavut.

[iv] The question included: Facebook, Instagram, LinkedIn, MeetUp, Pinterest, Reddit, Snapchat, Tumblr, Twitter, and YouTube.

[v] Small changes in R^2 values in social science research are common—even when the initial R^2 are considered small. Some examples of studies published in peer-reviewed journals include: Gearhart & Zhang, 2014; Johnson & Kaye, 2015; Lee, Chung & Park, 2016.