Statistically Evaluating Social Media Sentiment Trends towards COVID-19 Non-Pharmaceutical Interventions with Event Studies

Jingcheng Niu[1,2,3] **Erin E. Rees**[1] **Victoria Ng**[1] **Gerald Penn**[2,3]
[1]Public Health Agency of Canada [2]University of Toronto [3]Vector Institute
{niu,gpenn}@cs.toronto.edu
{erin.rees,victoria.ng}@canada.ca

Abstract

In the midst of a global pandemic, understanding the public's opinion of their government's policy-level, non-pharmaceutical interventions (NPIs) is a crucial component of the health-policy-making process. Prior work on COVID-19 NPI sentiment analysis by the epidemiological community has proceeded without a method for properly attributing sentiment changes to events, an ability to distinguish the influence of various events across time, a coherent model for predicting the public's opinion of future events of the same sort, nor even a means of conducting significance tests. We argue here that this urgently needed evaluation method does already exist. In the financial sector, *event studies* of the fluctuations in a publicly traded company's stock price are commonplace for determining the effects of earnings announcements, product placements, etc. The same method is suitable for analysing temporal sentiment variation in the light of policy-level NPIs. We provide a case study of Twitter sentiment towards policy-level NPIs in Canada. Our results confirm a generally positive connection between the announcements of NPIs and Twitter sentiment, and we document a promising correlation between the results of this study and a public-health survey of popular compliance with NPIs.

1 Introduction

As COVID-19 spreads rapidly around the world, governments have implemented different NPIs to contain the spread of the virus. While effective at slowing down the spread of COVID-19 (Haug et al., 2020), NPIs such as school and non-essential businesses closures, telecommuting, mask requirements and physical distancing measures have drastically changed our lives and sparked dissent. Anti-mask and anti-lockdown protests are commonplace, while there are nearly fifty million active cases around the world. It is crucial for decision makers to understand the public's opinion about NPIs, and for policy-makers to have a means of forecasting the level of popular compliance with them. This will determine their effectiveness as well as whether additional measures and communication strategies are needed in light of waning adherence.

Analysis of social media data is already popular among epidemiologists, as it is a data source with near real-time feedback at very low cost (Majumder et al., 2016). Extracting sentiment trends towards the pandemic on various social media platforms has already attracted interest (Wang et al., 2020b; Li et al., 2020; Wang et al., 2020a). Neural sentiment analysis is very prevalent because of its high performance on classification tasks[1] and versatility. Temporal variation of sentiment is usually represented by time series, in which an average model-predicted sentiment scores over from all social media posts within each time interval is computed. Previous work following this paradigm suffers from two major issues, however.

Firstly, nearly all time-series analyses have been based on sentiment classification results — every post is classified into one of the predetermined sentiment categories (positive/(neutral)/negative) — even though sentiment is a continuous random variable. For example, Wang et al. (2020b) provide two "sentiment-neutral" examples that in fact have differing sentiments. Smoothing sentiment from a continuous variable into a ternary or binary scale causes a loss of dynamics, hence increasing the difficulty of the task and lowering the reliability of all subsequent analyses. There are now n-valued sentiment corpora for $n = 5$ (Socher et al., 2013) and $n = 7$ (Mohammad et al., 2018), but finer-grained discrete sentiment does not entirely solve the problem. The valence regression task (*V-reg*) proposed by Mohammad et al. (2018) is far more suitable because it conveys a continuous sentiment intensity measure through a logistic regression score.

[1]Top performers achieve near perfect accuracies, e.g., Jiang et al. (2020) at 97.5%.

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 1–6
June 10, 2021.

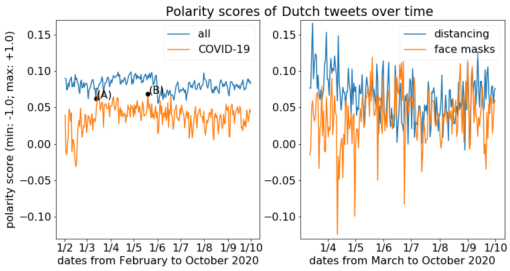


Figure 1: Wang et al. (2020a) claimed that general sentiment reached a minimum when the government announced a "lock-down" (A), and COVID-19 related sentiment reached a maximum when Amsterdam announced release measures (B). Note that the magnitude of difference between the minimum point they discovered at (A) and the valley a few days prior, at which there was no press conference, is not visible to the naked eye.

A continuous score also allows us to compute an average sample sentiment over a definite period of time, which has a more accurate variance than smoothing binary scores.

Secondly, because of the community's lack of a model capable of conducting significance tests and distinguishing the influence of various events across time, no statistically sound conclusion can be drawn. As an example, Wang et al. (2020a) claimed to have noticed a link between public sentiment and the timing of the Dutch government's press conferences by visually inspecting the raw trend of social media sentiment, seen in Figure 1. In fact, there were numerous peaks and valleys throughout the interval they studied, because the average sentiment fluctuated wildly during this time.

We can bring the potential of this urgently needed application to fruition by looking outside CL/NLP. Financial analysts face similar problems when they try to assess the effect of a particular news event on the price of a particular stock, because the price is affected by countless events as well as the reactions of traders with different motivations and perspectives on those events. *Event studies* (Brown and Warner, 1980, 1985) have been proposed and recognised as viable methods for attributing stock price fluctuations to specific financial events. To our knowledge, there has been no study of this class of methods within epidemiology.

2 Event Attribution

2.1 In Finance

In the financial sector, event studies are used to examine the return behaviour of a security after the market experiences some event (e.g., a stock split or an earnings release) that pertains to the firm that issued the security. The actual return of a stock (or a portfolio of assets) (R_t) at a given time t ($t = 0$ represents the time of the event) can be decomposed as follows: $R_t = \mathbb{E}[R_t|X_t] + \xi_t$. $\mathbb{E}[R_t|X_t]$ is an expected return, which can be explained by a model given the conditioning information X_t. ξ_t is an "abnormal" return that directly measures the unexpected changes on the returns, which are likely to have been caused by some unforeseen event (Eckbo, 2009). It is also possible that the abnormal return was just caused by chance ($\mathbb{E}[\xi_t] = 0$), however, and we can measure the statistical significance with which we can reject this null hypothesis through various tests based upon *time-series aggregation*, which we discuss presently.

The expected return can be estimated by a *market model* (Fama and MacBeth, 1973): $\mathbb{E}[R_t|X_t] = \alpha + \beta R_{m,t}$, where $R_{m,t}$ is the return of a market portfolio, i.e., of all of the assets in the market as represented by a broad market index (e.g., S&P 500, Nasdaq). β is the risk factor of the stock and can be computed using the ratio of the covariance between the actual return and the market return to the variance of the market return $\beta = \frac{cov(R,R_m)}{\sigma^2(R_m)}$. α is the bias that can be computed with least squares estimation, but since β is already computed, the optimal value of α is $\frac{1}{N}\sum_t^N (R_t - \beta R_{m,t})$ where N is the sample size.

The analysis of an event proceeds by first determining whether there is a statistically significant impact, and then if there is, computing the magnitude of the impact. To answer these two questions, the integral of the abnormal return, called the *cumulative average residual (CAR)*, is computed: $\text{CAR}(t_1, t_2) = \sum_{t=t_1}^{t_2} \xi_t$. Under the assumption that the return of a stock with no marked events is a stochastic process that perfectly reflects the overall performance of the market as accounted for by the market model (Fama and MacBeth, 1973), the expectation of CAR should be zero. Thus, we can test the null hypothesis that the event has no impact on the return, $\mathbb{E}[\xi_t] = 0$, by a one-sample t-test, one-sample Wilcoxon signed rank test (Wilcoxon, 1945), or a binomial proportionality z-test. In finance, the ratio of CAR divided by the overall actual return is traditionally used to represent the magnitude of an event's impact, but the statistics of these tests can also be used.

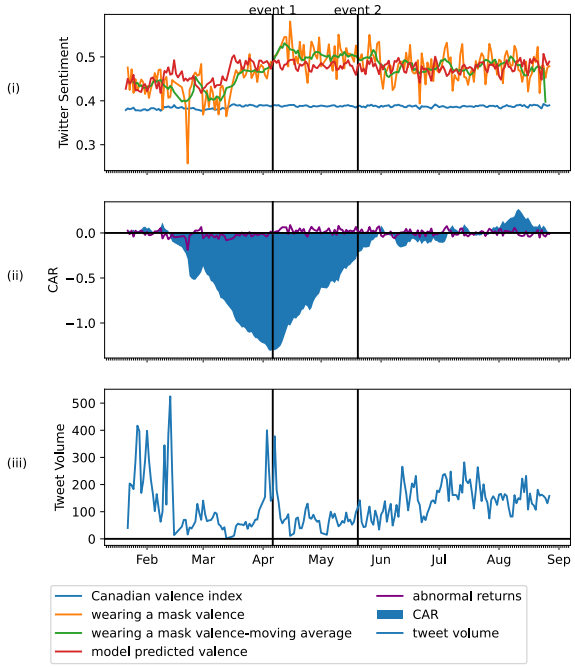
2.2 In Public Health

Over the course of the pandemic, governments around the world have utilized different NPIs at different times and with different stringencies (Hale et al., 2020). Therefore, overall sentiment shift cannot represent the impact of individual public health events. Instead, overall sentiment acts like market return: an aggregation of individual sentiments. Therefore, we define the daily sentiment index (I) as the average sentiment (valence) of all the tweets from a single day. Individual COVID-19-related topics are analogous to individual stocks, and the sentiment change on individual topics is reflected in the change of the sentiment index. But some topics specifically relate to certain events, similar to how individual stocks react to the news relevant to their firms. Therefore, the average sentiment $S_{m,t}$ of all discussions on topic m at time t is similar to the return of a stock in the event study. Our "market model" for sentiment is: $\mathbb{E}[S_{m,t}] = \alpha_m + \beta_m I_t$. We compute the abnormal sentiment by $\xi_{m,t} = S_{m,t} - \mathbb{E}[S_{m,t}]$ and calculate CAR by aggregating $\xi_{m,t}$ over time: $\text{CAR}(t_1, t_2) = \sum_{t=t_1}^{t_2} \xi_t$.

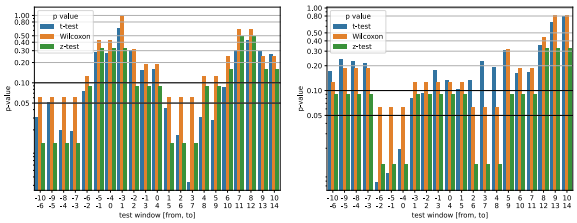
3 Experimental Setup

Gilbert et al. (2020) started collecting COVID-19 related tweets by searching for tweets mentioning at least one of the various naming conventions for COVID-19 using the Twitter search API as at January 21, 2020, and collected 281,487,148 tweets up until August 23rd, 2021. After Carmen geolocation (Dredze et al., 2013), we obtained 5,979,759 English Twitter samples from Canada.

For this paper, we studied two NPIs: wearing a mask and social distancing. For present purposes, we considered an event to be every change in the stringency level of any NPI, as measured by the Oxford COVID-19 Government Response Tracker (OxCGRT) project (Hale et al., 2020). We used a keyword-based filter to obtain topic-related tweets. We began with a manually written list of related keywords to obtain a list of tweets M that contain a keyword, and $\overline{M}$ that do not contain any keyword. Then for each bigram and trigram x, we calculated a topic relevance score based on pointwise mutual information: $pmi(x; M) - pmi(x; \overline{M})$. We ranked the top 150 keywords for each n-gram and manually removed the topic-unrelated ones. For example, "covidsafe" was identified using this method but "congressman sponsor," a topic relevance score of 14.59, was nevertheless manually removed.



(a) Wearing a mask sentiment analysis



(b) Event 1 significance study (c) Event 2 significance study

Figure 2: Wearing a mask event significance

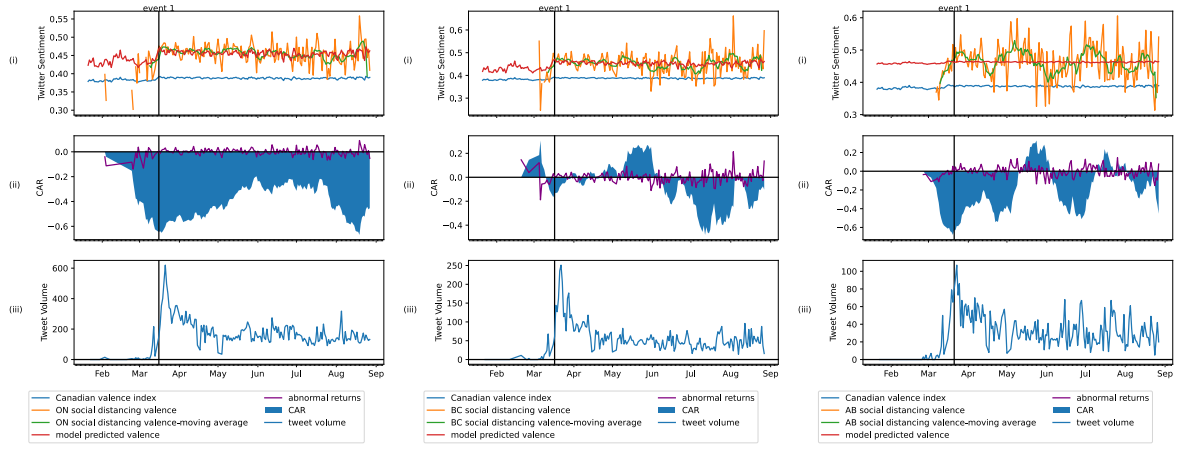
After filtering all the tweets connected to an NPI of interest, we computed their valence score using the NTUA-SLP model,[2] which was selected from the 75 entrants to the V-reg shared task (Mohammad et al., 2018). We followed the hyperparameter settings from the original paper (Baziotis et al., 2018) and reproduced its reported Pearson correlation (0.846) on the English valence dataset. To establish a periodic time series of valence change, we computed the daily average valence of tweets posted on the same day.[3]

4 Individual NPIs Experimental Results

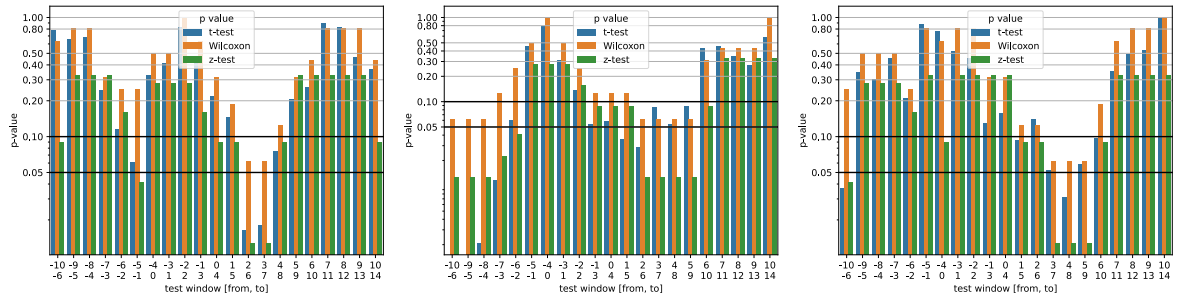
Wearing A Mask Canada's mask advisory has changed several times during the progression of the pandemic (Mohammed et al., 2020) and we investigated two key changing points of the advisory

[2] https://github.com/cbaziotis/ntua-slp-semeval2018

[3] Our subsequent analyses and data are publicly available: https://github.com/frankniujc/covid_sentiment_analysis.



(a) Ontario (initial: Mar 16) (b) British Columbia (initial: Mar 17) (c) Alberta (initial: Mar 21)



(d) ON significance (e) BC significance (f) AB significance

Figure 3: Social distancing recommendation event significance by province.

as events. On April 6th, 2020, the Public Health Agency of Canada (PHAC) revised the advisory for mask wearing (event 1), permitting the use of non-medical face coverings in public (Chase, 2020; Mohammed et al., 2020). Finally on May 20, 2020, PHAC formally issued a recommendation for the general public to wear masks in public (event 2) (Mohammed et al., 2020; Harris, 2020).

Assuming a confidence threshold of $\alpha = 0.05$, event 1 had a statistically significant positive impact for up to 9 days (Figure 2b). Event 2 also showed significance from two days after the event to up to eight days after ($[+2, +8]$; Figure 2c). Unlike event 1, there is also a period of significance right before the event occurred. This may have been anticipatory, or it may indicate that the observed impact had instead been caused by prior events. During the 9-day effect window of event 1, there is a 2.13% positive CAR, with t-statistic 1.73, Wilcoxon statistic 7.0, and z-statistic 1.67.

Social Distancing Social distancing recommendations have been issued with different stringencies and at different times at the provincial level in Canada. Therefore, we focus separately on three provinces: Ontario (ON), British Columbia (BC) and Alberta (AB), with sufficient numbers of tweets and different distancing policies. According to McCoy et al. (2020), Ontario released its first province-wide social distancing recommendation on March 16, 2020 (Williams, 2020); British Columbia issued a social distancing recommendation on March 17, 2020 (Dix and Henry, 2020); and lastly, Alberta released a public message about social distancing on March 21st[4] (McCoy et al., 2020).

Figure 3 analyses the significance of the initial recommendations in those three provinces. All three announcements have a positive impact on CAR with statistical significance. Ontario's recommendation (Figure 3d) has a short but significant impact on $[+2, +7]$. Alberta (Figure 3f) exhibits a significant impact on $[+3, +9]$, and British Columbia on $[+1, +9]$.

5 CAR and Survey Data Correlation

To help understand whether the sentiment of NPIs measured using Twitter are representative of the general Canadian population, we assessed the correlation between our NPI sentiments and the level of compliance measured through a national survey.

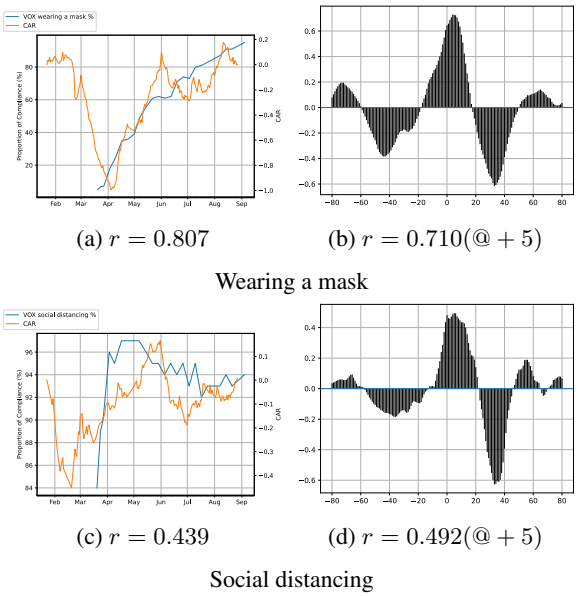
[4] https://www.alberta.ca/prevent-the-spread.aspx

The COVID-19 Monitor initiative (COV, 2020; Mohammed et al., 2020) has conducted 25 surveys in Canada on people's compliance with 6 NPIs since mid-March. Each survey has approximately 2000 participants. The demographics of the participants have been pre-stratified, and each wave was post-stratified by modelling raking weights based on the 2010 Canadian Census. Among the 6 NPIs, both *social distancing* and *wearing a mask* appear. For the cross-correlation test, both time series have been detrended using the SciPy signal package,[5] and then pre-whitened following the instructions proposed by Dean and Dunsmuir (2016) to remove autocorrelations with the time series.[6] Figure 4 shows the correlations and cross-correlations with the proportion of the population who report complying with either of these two NPIs and CAR. Wearing a mask receives a strong Pearson $r = 0.915$ (Figure 4a), a cross-correlation of 0.710 and a +5 lag, meaning CAR is 5 days ahead of the survey (Figure 4b). Social distancing receives a moderate Pearson $r = 0.481$ (Figure 4c), a cross-correlation of 0.492 and also a +5 lag (Figure 4d). The cross-correlations cannot be quantitatively compared with the Pearson correlation scores as they are calculated differently, but the general trend stays the same: wearing a mask exhibits a strong correlation while social distancing, only moderate one. The lags also accord with our expectations as COV (2020) conducted surveys 4 to 10 days apart.

The lower correlation for social distancing might have been caused by their more diverse implementation across subsovereign jurisdictions (see section 4). As the details of the sample selection process at the provincial level are not publicly available, we have not been able to draw direct, provincial comparisons. Mask-wearing advisories, however, are mostly issued at the federal level in Canada. Comparing mask-wearing across provinces is thus less problematic. With both types of NPI, Twitter users are demographically younger, better educated, and more urban than the general population (Mellon and Prosser, 2017; Murthy et al., 2016). This may explain some differences from the national distribution sampled for this survey.

[5] https://docs.scipy.org/doc/scipy/reference/signal.html

[6] We tested the autocorrelation of both the CARs and survey data. The level of autocorrelation in all the time series is low, and applying pre-whitening did not result in different conclusions in this study.



(a) $r = 0.807$ (b) $r = 0.710(@+5)$

Wearing a mask

(c) $r = 0.439$ (d) $r = 0.492(@+5)$

Social distancing

Figure 4: CAR and compliance survey correlation. Captions of (a) and (c) report Pearson correlations; captions of (b) and (d) report cross-correlations with days of lag.

Acknowledgements

This study is funded by the Canadian Safety and Security Program (CSSP) from Defence Research and Development Canada (DRDC). Many thanks for technical support and study design guidance to Jean-Philippe Gilbert (Laval University), Hélène Carabin, Esther Perez, Mireille D'Astous, Simon de Montigny, and Nasri Bouchra (University of Montreal), Patrick Daley (Public Health Agency of Canada), and Suzanne Hindmarch (University of New Brunswick). We also thank Saif Mohammad (National Research Council of Canada) for his help with sentiment analysis, and Tong Lin (University of Toronto) for his help with event studies.

References

2020. COVID-19 Monitor. Technical report, Vox Pop Labs.

C. Baziotis, A. Nikolaos, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 245–255.

S. Brown and J. Warner. 1985. Using daily stock returns. *Journal of Financial Economics*, 14(1):3–31.

S. J. Brown and J. B. Warner. 1980. Measuring security price performance. *Journal of Financial Economics*, 8(3):205–258.

S. Chase. 2020. Theresa Tam offers new advice: Wear a non-medical face mask when shopping or using public transit. *CBC News*.

R. T. Dean and W. T. M. Dunsmuir. 2016. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods*, 48(2):783–802.

A. Dix and B. Henry. 2020. Joint statement on Province of B.C.'s COVID-19 response, latest updates | BC Gov News. *BC Gov News*.

M. Dredze, M. J. Paul, S. Bergsma, and H. Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

B. E. Eckbo, ed. 2009. *Handbook of Corporate Finance: Empirical Corporate Finance. Vol. 1: ...*, reprinted edition. Number 3 in Handbooks in Finance. Elsevier North-Holland, Amsterdam.

E. F. Fama and J. D. MacBeth. 1973. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81(3):607–636.

J.-P. Gilbert, J. Niu, S. de Montigny, V. Ng, and E. E. Rees. 2020. Machine Learning Identification of Self-Reported COVID-19 Symptoms from Tweets in Canada. In *AAAI 2021 W3PHIAI-21 Workshop*.

T. Hale, N. Angrist, E. Cameron-Blake, L. Hallas, B. Kira, S. Majumdar, A. Petherick, T. Phillips, H. Tatlow, and S. Webster. 2020. Variation in government responses to COVID-19.

K. Harris. 2020. Canadians should wear masks as an 'added layer of protection,' says Tam. *CBC News*.

N. Haug, L. Geyrhofer, A. Londei, E. Dervic, A. Desvars-Larrive, V. Loreto, B. Pinior, S. Thurner, and P. Klimek. 2020. Ranking the effectiveness of worldwide COVID-19 government interventions. *medRxiv* 2020.07.06.20147199.

H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao. 2020. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu. 2020. The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. *International Journal of Environmental Research and Public Health*, 17(6):2032.

M. S. Majumder, M. Santillana, S. R. Mekaru, D. P. McGinnis, K. Khan, and J. S. Brownstein. 2016. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR Public Health and Surveillance*, 2(1):e30.

L. G. McCoy, J. Smith, K. Anchuri, I. Berry, J. Pineda, V. Harish, A. T. Lam, S. E. Yi, S. Hu, L. Rosella, B. Fine. 2020. Characterizing early Canadian federal, provincial, territorial and municipal nonpharmaceutical interventions in response to COVID-19: A descriptive analysis. *CMAJ Open*, 8(3):E545–E553.

J. Mellon and C. Prosser. 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3):2053168017720008.

S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

A. Mohammed, R. M. Johnston, and C. van der Linden. 2020. Public Responses to Policy Reversals: The Case of Mask Usage in Canada during COVID-19. *Canadian Public Policy*, 46(S2):S119–S126.

D. Murthy, A. Gross, and A. Pensavalle. 2016. Urban Social Media Demographics: An Exploration of Twitter Use in Major American Cities. *Journal of Computer-Mediated Communication*, 21(1):33–49.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

S. Wang, M. Schraagen, E. T. Kim Sang, and M. Dastani. 2020a. Public Sentiment on Governmental COVID-19 Measures in Dutch Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

T. Wang, K. Lu, K. P. Chow, and Q. Zhu. 2020b. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access*, 8:138162–138169.

F. Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

D. Williams. 2020. Enhanced Measures to Protect Ontarians from COVID-19. *Ontario Newsroom*.