Words Speak Louder Than Numbers

Estimating China's COVID-19 Severity with Deep Learning

Julian TszKin Chan, Kwan-Yuet Ho, Kit Lee, Kawai Leung, and Weifeng Zhong

WORKING PAPER

COVID-19 RESPONSE

MERCATUS CENTER
George Mason University
3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

Julian TszKin Chan et al. "Words Speak Louder Than Numbers: Estimating China's COVID-19 Severity with Deep Learning." Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, December 2020.

Abstract

We develop a deep learning algorithm to estimate the severity of the 2020 COVID-19 outbreak in China by analyzing the language of the *People's Daily*, China's official newspaper. The algorithm uses the 2002–2003 SARS outbreak as the benchmark and learns how the newspaper's language evolved during the epidemic cycle. It then maps the daily coverage of the coronavirus outbreak to the SARS timeline and, hence, estimates its relative position in the benchmark epidemic cycle. We call this timeline-based measure the Policy Change Index for Outbreak. We find a pronounced discrepancy between our severity measure and China's official numbers of diagnosed cases. We also demonstrate that our indicator is more informative about the outbreak's severity than a conventional sentiment analysis.

JEL codes: C63, D83, I18, P49

Keywords: policy change, propaganda, deep learning, coronavirus, COVID-19, SARS, outbreak

Author Affiliation and Contact Information

Julian TszKin Chan
Managing Economist and Lead Data Scientist
Bates White Economic Consulting
julian.chan@policychangeindex.org

Kwan-Yuet Ho
Data Scientist
Leidos
KwanYuet.Ho@leidos.com

Kit Lee
Data Scientist

Kawai Leung
Independent Researcher

Weifeng Zhong (corresponding author)
Senior Research Fellow
Mercatus Center at George Mason University
wzhong@mercatus.gmu.edu

Acknowledgments

We appreciate helpful feedback from King-Wah Fu, Adrienne Keen, Walter Valdivia, and participants at the Association of Chinese Political Studies annual meeting, at the Data Science Conference on COVID-19, and at the NLP Summit. We thank NVIDIA for the generous donation of a TITAN Xp, which was used for this research. The views expressed here are solely the authors' and do not represent the views of Bates White, Leidos, the Mercatus Center, or their other employees.

This paper can be accessed at https://www.mercatus.org/publications/covid-19-crisis-response/words-speak-louder-numbers.

Words Speak Louder Than Numbers:

Estimating China's COVID-19 Severity with Deep Learning

Julian TszKin Chan, Kwan-Yuet Ho, Kit Lee, Kawai Leung, and Weifeng Zhong

1. Introduction

How severe was coronavirus disease 2019 (COVID-19) in China, really? It is widely suspected, both inside and outside China, that the country's official numbers of diagnosed cases understate the extent of the outbreak. Even the official statistics themselves are incoherent. On February 13, 2020, the Chinese authorities confirmed more than 15,000 new cases nationwide—40 times the previous day's number—because of a change in counting criteria. Just two months after that, they revised the death toll for the city of Wuhan, the epicenter, upward by 50 percent, citing unspecified omissions.

The COVID-19 pandemic is a sobering reminder that the time when China's opacity had few implications for other countries has passed. As China's global influence grows and the US-China rivalry intensifies, policymakers' need to have a good grasp on current events in China is stronger than ever. While the actual scale of the disease in China may remain a mystery, any step closer to the truth would enhance our understanding of the Chinese government's behavior for years to come.

In this paper, we measure the severity of COVID-19 in China, not through the Chinese government's official numbers but through how state-controlled media talked about the outbreak. Words can speak louder than (some) numbers. While it may be simple to release false statistics, it is more difficult to conceal the truth when the government has to address a crisis, such as a severe infectious disease, at length in the media. The discrepancy between what the government says and what numbers are released, if any, would be indicative of the level of underreporting in the official statistics. For example, on January 23, 2020, the Chinese government announced a lockdown of Wuhan, a city with a population of 11 million, and its neighboring cities and discussed the necessity of doing so in state media. However, up to that day, the Chinese authorities had confirmed a total of fewer than 600 cases of the novel coronavirus across the country. Therefore, changes in words during the outbreak may provide us with a clearer picture of the severity than official numbers do.[1]

Building on previous work by Chan and Zhong (2019, 2020), we develop a deep learning algorithm to measure changes in language surrounding COVID-19. We use the 2002–2003 severe acute respiratory syndrome (SARS) outbreak as the benchmark, and we use the words in the *People's Daily*—China's most prominent official newspaper, in Chinese—as input data.[2]

It is now known that the major part of the SARS outbreak spanned from November 2002 to July 2003. The language the *People's Daily* used to describe it evolved during that time too. When the *People's Daily* first reported the outbreak on February 12, 2003, it said there was no need to be concerned as long as the public took appropriate measures. However, on May 11, around the time when the outbreak was the most severe, it called for officials and residents in Beijing to make every effort possible to contain the disease's spread. After SARS hit a plateau, victory-claiming language became increasingly frequent in the newspaper.

Using the SARS cycle as the precedent, we map the current daily coverage of COVID-19 to that of SARS. As described in detail in section 4, we build a natural language processing

---

[1] Even Chinese officials are known to have relied on alternatives to their official numbers for policymaking. Premier Li Keqiang, for example, once admitted that GDP figures in China were unreliable and that he instead used three other indicators: railway freight, electricity consumption, and bank loans. The composition of the three was later termed the "Li Keqiang Index" (*The Economist* 2010).

[2] The source code of the project can be found at https://github.com/PSLmodels/PCI-Outbreak, and subsequent updates of the indicator are available at "Policy Change Index (PCI)," https://policychangeindex.org.

algorithm that learns where SARS articles fall in the SARS timeline and, subsequently, assesses when COVID-19 articles were published. Because the model is built to understand the tone and tenor of SARS-episode language, it will likely place COVID-19 articles in the SARS era, too. If the model places a COVID-19 article at the peak time during the SARS outbreak—a placement that would be factually incorrect because COVID-19 articles appeared in the newspaper 17 years later—this will indicate that the article appeared when the COVID-19 outbreak was also near its own peak. We call this measure of severity the Policy Change Index for Outbreak (PCI-Outbreak).

Several results, discussed in more detail in section 5, follow from the contrast between the PCI-Outbreak and the official numbers. First, both reached their peaks in February 2020, suggesting that a discrepancy between the two measures was not present during the first part of the epidemic cycle. Second, after the outbreak plateaued in February, the improvement measured by the PCI-Outbreak is much slower than the improvement reported by the official numbers. The increasing discrepancy between the two measures, therefore, suggests that the Chinese authorities may have under-reported the cases, as illustrated by the tone and tenor of their own language. Finally, when resurgences of cases occurred, such as the second wave in Beijing in June and a new outbreak in Xinjiang in July and August, these events were more severe than the official numbers suggest, according to our measure.

Granted, the true scale of the outbreak may never be known. The discrepancy between the official numbers and the Chinese government's own words is nevertheless an indication of possible under-reporting in the statistics. In section 6, we compare the PCI-Outbreak findings to a conventional sentiment analysis, and we provide robustness checks on our main results.

2. Literature Review

In gauging the actual severity of COVID-19 in China, two approaches have emerged in the literature. Researchers who adopt the first approach ask the following question: Had the official statistics been coherent over time—even if they were understating the truth—how many infections would there have been? Li et al. (2020) estimate the prevalence of undocumented cases (due to lack of symptoms, for example) that would have been consistent with China's officially confirmed cases. They find that before the January 23, 2020, lockdown of Wuhan, the number of undocumented infections could have been six times the number of confirmed ones. A similar study by Zhao et al. (2020) suggests that the scale of undocumented cases may have been even larger at the earlier stage of the Wuhan outbreak. Tsang et al. (2020) focus on a different aspect of the statistics: the frequently changing definitions according to which Chinese authorities confirmed cases. The researchers estimate that had the widest definition—out of the seven the Chinese authorities had used over time—been applied consistently throughout the outbreak, there would have been four times as many confirmed cases by February 20, 2020.

Researchers who adopt the second approach investigate the validity of China's official numbers. Leveraging travel data, Wu, Leung, and Leung (2020) estimate the size of the outbreak in Wuhan on the basis of reported Wuhan-originating cases outside mainland China. They find that, to be consistent with confirmed cases overseas, the number for Wuhan as of January 25, 2020, would have had to be more than 100 times larger than what the authorities confirmed. The back-of-the-envelope analyses by Imai et al. (2020a, 2020b) and Scissors (2020), using similar approaches, also show a large discrepancy between the official numbers and the possible true scale of the outbreak.

Our paper complements the above literature in that, instead of estimating the case numbers from an epidemiological standpoint, we deduce the severity from the words of the Chinese government—which presumably has more information about the outbreak than the public does.

Our paper joins a growing literature that makes policy inferences from analyzing propaganda. Pioneered by the work by Lasswell et al. (1949) and George (1959), inductive analysis of propaganda messages has been a standard methodology among academics and policy practitioners.[3] But for a long time this technique was labor-intensive because, like any other text, propaganda messages are unstructured data.

In recent years, however, advances in natural language processing and deep learning have led to a boom in the literature by making analyses of unstructured data more accessible. In the area of monetary policy, Zirn, Meusel, and Stuckenschmidt (2015) study the formation of opinion groups in the Federal Open Market Committee discussions. Guo et al. (2015) and Schonhardt-Bailey (2013) explore the influence of members of the Federal Open Market Committee on one another. Other studies apply similar techniques to the prediction of lawmaking, such as predicting whether a US congressional bill will survive the committee process (Yano, Smith, and Wilkerson 2012) or whether a bill will be voted and enacted into law (Gerrish and Blei 2011; Kraft, Jain, and Rush 2016; Nay 2017). Predictions of court rulings are also explored for the US Supreme Court (Agrawal et al. 2017; Katz, Bommarito, and Blackman 2017; Sim, Routledge, and Smith 2016), the German Fiscal Courts (Waltl et al. 2017), and the European Court of Human Rights (Aletras et al. 2016). Finally, Chan and Zhong (2019, 2020)

[3] See Krippendorff and Bock (2009) and the references therein for this tradition in the literature.

use natural language processing to predict Chinese government actions on the basis of *People's Daily* text.

3. Data

Our methodology requires that the text from which we make inferences be generated in a controlled messaging environment, such as disciplined state-owned media. In this section, we describe our data source, the *People's Daily*, and discuss why it is an appropriate input to our model.

3.1. Why the People's Daily

Just like *Pravda* for the Soviet Union, the *People's Daily* has been the most prominent official newspaper throughout the history of the People's Republic of China. Government propagandists use it to promote official viewpoints and communicate top-down directives to the rest of the country.

The importance of this newspaper is most apparent from its organizational connection to the Communist Party of China. The *People's Daily* is similar to the Soviet model in that it is the official publication of the party's central committee and is managed by its Department of Publicity. Historically, the editor in chief of the newspaper has often been concurrently a deputy director of the Department of Publicity as well.[4]

The central government of China has been directly involved in the publication of the *People's Daily*. According to Wu (1994), top officials of the Communist Party often command

[4] The editor in chief as of December 2020, Tuo Zhen, was a deputy director of the Department of Publicity before assuming the editorship.

subjects and specify constraints for the newspaper's editorial team. Editorials and commentators' articles are also subject to tight censorship by top officials before publication. The most important editorials, at times, must be approved by the general secretary of the Communist Party—that is, the Chinese president.

One may wonder how, since China's economic reform started in 1978, the official status of the *People's Daily* has evolved. There is no question that, in general, the media industry has become more commercialized in post-reform China. However, not only has commercialization not led to more diverse political voices in China (Brady 2008; Stockmann 2013), but the *People's Daily* has remained the least commercialized newspaper in the country, strictly upholding the party line in its reporting.[5] Previous work by Chan and Zhong (2019, 2020) on the Policy Change Index also suggests that the authoritative nature of this newspaper has remained largely unchanged throughout its existence.

3.2. Text and Epidemic Data

We have collected the full text of *People's Daily* articles published during two episodes: SARS and COVID-19. The SARS episode, which serves as our benchmark, begins on April 3, 2003, when the *People's Daily* started to cover the outbreak regularly, and ends on July 5, 2003, when the World Health Organization declared that the global SARS outbreak had been contained.[6] The COVID-19 episode begins on January 21, 2020, when the *People's Daily* started to cover the outbreak regularly, and ends on September 15, 2020, the time of this analysis.

[5] According to one statistic (Esarey 2006), from the 1980s to 2004, market competition from other, more commercialized media outlets caused the hard-copy circulation of the *People's Daily* to plummet from 5 million copies per day to below 2 million.
[6] The newspaper started to cover the SARS outbreak in January 2003, but the coverage before April 2003 was sporadic and, hence, not included in our analysis. Also not included in our analysis is the second SARS outbreak in China, which started in April 2004 and is separate from the benchmark case.

For the SARS episode, we consider two sets of articles: those that are relevant to the outbreak and those that are irrelevant to it. Relevant articles are those that mention SARS-related keywords, such as "atypical," "pneumonia," "SARS," and "epidemic" (translated from Chinese), frequently enough, while irrelevant articles are those that do not mention the keywords at all.[7] The two sets together allow our model to learn whether certain content is outbreak-related, and the first set allows the model to learn, conditionally on the article's being relevant, its time of publication. For the COVID-19 episode, we select articles that mention COVID-19-related keywords (such as "novel coronavirus," "pneumonia," "COVID," and "epidemic") frequently enough.[8]

In our corpus, each article is associated with its date of publication, title, body, and a binary indicator for whether the article is considered outbreak-related. Table 1 shows some summary statistics for each set of articles.

Table 1. Summary Statistics

	Number of articles	Average number of characters per article	Average number of sentences per article
SARS episode (relevant)	2,285	864	21
SARS episode (irrelevant)	8,577	639	17
COVID-19 episode (relevant)	5,672	1,361	33

Because the language representation we use (described in the next section) requires intensive use of computational resources, we segment each article into sentences and treat the sentence as the unit of observation for the model. We recognize that the sentence-level analysis loses some information, because multiple sentences, in a certain order, are also part of the

[7] We consider an article relevant if it mentions the keywords at least three times and at least three times per thousand characters.
[8] The frequency threshold is the same as for the SARS episode.

linguistic context. We take this sentence-level analysis as an exploratory step, and we leave the alternative approaches that have different units of observation, such as paragraph and article, to future studies.

As for the epidemic data officially disclosed by the Chinese authorities, we have collected, for each outbreak, the time series for the number of unresolved diagnosed cases, which is defined as the number of diagnosed cases where the person has neither recovered nor died. The SARS data are from the World Health Organization, while the COVID-19 data are from the Coronavirus Resource Center at Johns Hopkins University.

Figure 1 and figure 2 plot, for each episode, the number of relevant articles and the number of unresolved diagnosed cases over time. In both cases, the quantity of coverage seems to be correlated with the official numbers, suggesting that, loosely speaking, the state media tend to talk more about the outbreak when it is more severe. A priori, more coverage does not necessarily mean that the tone and tenor of the coverage are more pessimistic. Therefore, understanding the severity embedded in the language requires a method to "read the tea leaves." We turn to this in the next section.

Figure 1. ***People's Daily*** **Coverage and Official Statistics, SARS**

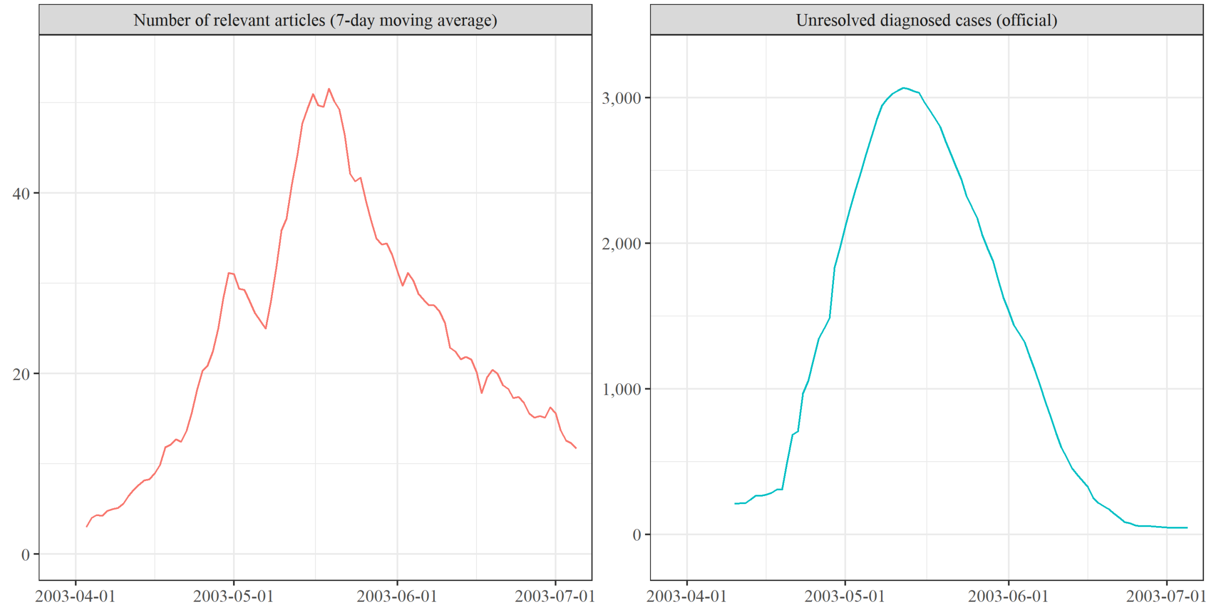
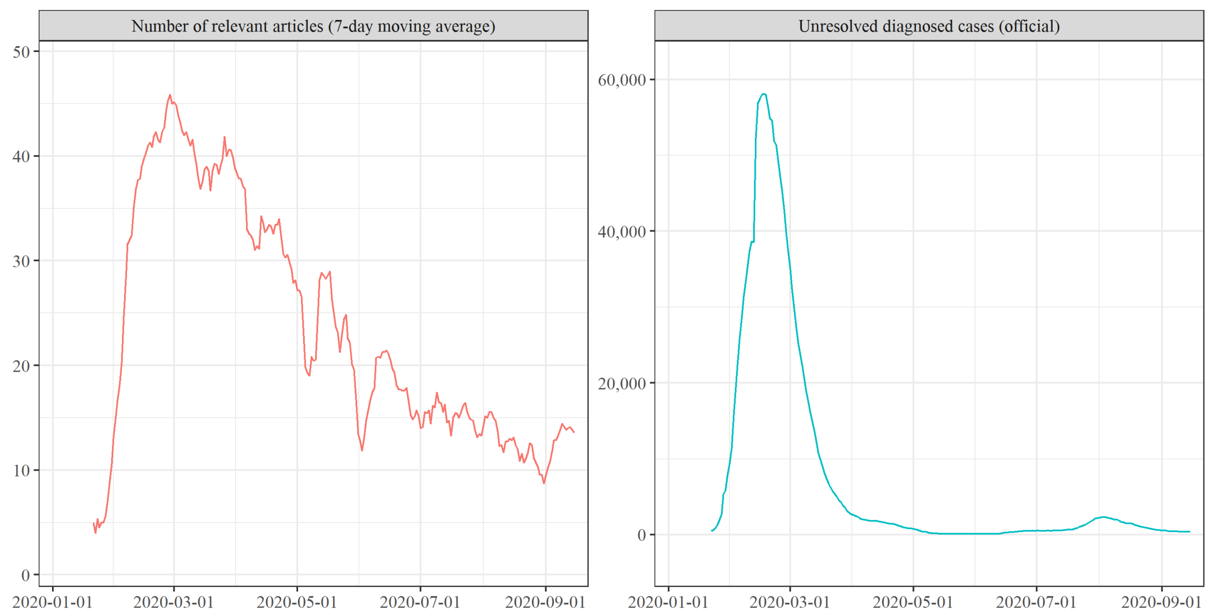


Figure 2. ***People's Daily*** **Coverage and Official Statistics, COVID-19**



4. Methodology

We describe below a deep learning algorithm on which the PCI-Outbreak is based. Using state-of-the-art techniques in natural language processing, we build a two-stage algorithm that predicts whether a piece of text is outbreak-related and, if so, when in the SARS news cycle such a text would have been published. We then convert the predicted time of publication of COVID-19-episode text to a relative index of severity on the basis of the SARS record of diagnosed cases.

Two fundamental assumptions underlie our approach. First, we assume that the Chinese authorities know more about the severity of an outbreak than the public does and that their choice of words in propaganda outlets at least in part reflects their private information. This assumption is plausible, given the opacity of the Chinese system and the crucial role propaganda plays in it. Second, we assume that the way the Chinese government handled propaganda during the SARS outbreak is largely being repeated during COVID-19. While the nature of rare events makes it difficult to validate the second assumption, previous research on the behavior of China's propaganda system suggests that it has been quite consistent over time.[9]

4.1. Pretrained Language Representation

Before building the two-stage algorithm, it is necessary to turn words, which are unstructured data (just like images and audio), to structured data that machine learning models can use as input. This step is called language representation.

There is a long history of practice in which machine learning models are trained to generate general representations (of natural languages), which can, in turn, be used to automate various language tasks, such as questions and answers, paraphrasing, translation, and

[9] See, for example, Brady (2008), Stockmann (2013), and Chan and Zhong (2019, 2020).

summarization. These models are called pretrained language representations. Popular ones include word2vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), and ELMo (Peters et al. 2018). This strategy was proved to be able to achieve high performance with lower computational cost compared to having to train a specific representation from scratch for each language task. In 2019, Google released the pretrained Bidirectional Encoder Representations from Transformers (BERT), which quickly became the gold standard for various language tasks (Devlin et al. 2019).

BERT has achieved state-of-the-art performance in question answering (Talmor et al. 2019), sentiment analysis (Xu et al. 2019), text classification (Kowsari et al. 2019), and other natural language processing tasks. Compared to static word embeddings such as word2vec, these improvements are a result of the contextual information encoded in BERT during pretraining. Because of the attention mechanism (Clark et al. 2019), BERT is able to generate word embeddings according to the preceding and following words in the same sentence or document. In this paper, we leverage this context-aware capability of BERT and modify it for our purposes.

4.2. Two-Stage Algorithm

Because BERT has a better performance on shorter texts and the sample size of SARS-related or COVID-19-related articles is small, we segment each article (its title and body) into sentences and treat the sentence as the unit of observation for the model. We do not, therefore, treat a sentence in the title differently from a sentence in the body of an article.

For a generic sentence i, let s_i denote the text of the sentence and t_i denote the time of the sentence's publication. We normalize both timelines such that time 0 represents the beginning of the episode's horizon.[10] Let r_i be a binary indicator, where $r_i = 1$ means the sentence is relevant to an epidemic outbreak at hand and $r_i = 0$ means it is not. We do not use any other characteristics of a sentence (such as the page where it appears or the author of the text) in our model.

The algorithm has two stages. In the first stage, the model predicts whether a sentence is relevant to the epidemic outbreak. Sentences predicted to be relevant advance to the second stage, in which the model predicts when in the SARS timeline the text would have been published. In each stage, we use stratified sampling (by date) to split the SARS-episode sample into training data and test data.[11]

For the first stage, we fine-tune a pretrained BERT model $BERT_1(\cdot)$ such that, for each sentence s_i, the algorithm gives a binary prediction $\hat{r}_i = BERT_1(s_i)$ for whether the sentence is relevant to the (SARS) outbreak. When fine-tuning the model, we formulate it as a binary classification problem and minimize the cross-entropy as the loss function.

For the second stage, we fine-tune another pretrained BERT model $BERT_2(\cdot)$ such that, for each sentence s_i, if the first-stage model deems it relevant—that is, if $BERT_1(s_i) = 1$—the algorithm gives a continuous prediction $\hat{t}_i = BERT_2(s_i)$ for when during the (SARS) outbreak the sentence was published. When we fine-tune the model, we formulate it as a regression problem and minimize the mean absolute error as the loss function.[12]

[10] Time 0 corresponds to April 3, 2003, in the SARS episode and January 21, 2020, in the COVID-19 episode.
[11] In each stage, 90 percent of the sentences are randomly assigned to the training set and the other 10 percent to the test set.
[12] That is, we treat the time of publication as a continuous variable that happens to have all integer realizations (i.e., in days) in the training data.

4.3. Constructing the PCI-Outbreak

To construct a severity measure, we deploy the two-stage algorithm $(BERT_1, BERT_2)$ to the COVID-19-episode text, and we map the algorithm's output, which consists of the predicted dates in the SARS timeline, to the normalized number of unresolved diagnosed SARS cases on a scale of 0 to 1. In other words, if the predicted date of a COVID-19-episode sentence is May 12, 2003, which is the day when the number of SARS cases reached the maximum (3,068 cases), then the predicted (normalized) severity score for that sentence is 1.[13]

Because multiple sentences are published each day, we take the mean of the severity scores for all sentences published on the same day as the (aggregated) severity score for that day.[14] We call this date-level severity score the PCI-Outbreak.

Note that, to map a date in the SARS timeline to a severity score, some proxy for the (unknown) reality of the SARS outbreak is needed. We use the official statistics about SARS as that proxy, though we are aware that the SARS numbers may have been deflated. But because the SARS outbreak was contained, the reported cases do show an evolution from emergence to peak to containment. Regardless of how deflated the SARS numbers are, if we observe a discrepancy between the official statistics about COVID-19 and the PCI-Outbreak, it may mean that the COVID-19 numbers were *even more deflated* than the SARS numbers.

4.4. Summary of Model Performance

In the training phase, the first-stage model has an accuracy rate of 97 percent; the accuracy rate declines to 90 percent in the testing phase. Table 2 shows the confusion matrix using the training

[13] Note that each severity score can correspond to two dates in the SARS timeline (one before the peak and the other after the peak), but each date in the SARS timeline has a unique severity score.
[14] We recognize that by taking the average we are ignoring how many sentences were published on each day. But the information on the frequency of mentions can be obtained from the summary statistics in figure 2.

and testing data. The results suggest that the increases in error rates are split evenly between false positives and false negatives.

Table 2. Confusion Matrix of the First-Stage Model

	Predicted value			
	Training		Testing	
True value	Not relevant	Relevant	Not relevant	Relevant
Not relevant	72%	2%	68%	6%
Relevant	0%	25%	4%	22%

The second-stage model predicts the publication dates of sentences during the SARS outbreak. We use mean absolute error (MAE) to assess the performance of the model. The MAEs are 2.4 and 11.7 during the training and testing phases, respectively.

Besides the MAE, we created a confusion matrix to visualize the performance of the second-stage model. We transform the actual and predicted dates into the actual and predicted weeks since May 12, 2003. Table 3 shows the confusion matrix of the second-stage model using the training data; table 4 shows the counterpart using the testing data. As in the first stage and as expected, the performance decreases from training to testing in the second stage as well.

Table 3. Confusion Matrix of the Second-Stage Model with Training Data

True week	Predicted week													
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1.51	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.84	1.61	0.03	0.02	0.01	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00
2	0.01	1.03	3.56	0.14	0.03	0.01	0.02	0.05	0.00	0.01	0.00	0.00	0.00	0.00
3	0.00	0.03	1.90	7.42	0.42	0.07	0.06	0.08	0.02	0.01	0.00	0.00	0.00	0.00
4	0.00	0.01	0.02	2.06	4.78	0.33	0.04	0.08	0.02	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.02	0.30	3.90	7.96	0.90	0.32	0.03	0.03	0.00	0.00	0.00	0.00
6	0.00	0.00	0.01	0.04	0.33	4.70	8.61	1.46	0.07	0.03	0.01	0.01	0.00	0.00
7	0.00	0.00	0.00	0.01	0.06	0.12	2.17	7.13	1.06	0.04	0.00	0.00	0.00	0.00
8	0.00	0.00	0.02	0.00	0.02	0.04	0.08	1.95	7.54	1.11	0.03	0.02	0.00	0.00
9	0.00	0.00	0.00	0.01	0.01	0.01	0.04	0.12	1.01	6.23	0.38	0.03	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.03	0.90	3.86	0.45	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.01	0.04	0.66	3.37	0.20	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.00	0.01	0.02	0.41	3.45	0.40
13	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.01	0.02	0.20	1.41

Note: The numbers are percentage points and they sum to 100.

Table 4. Confusion Matrix of the Second-Stage Model with Testing Data

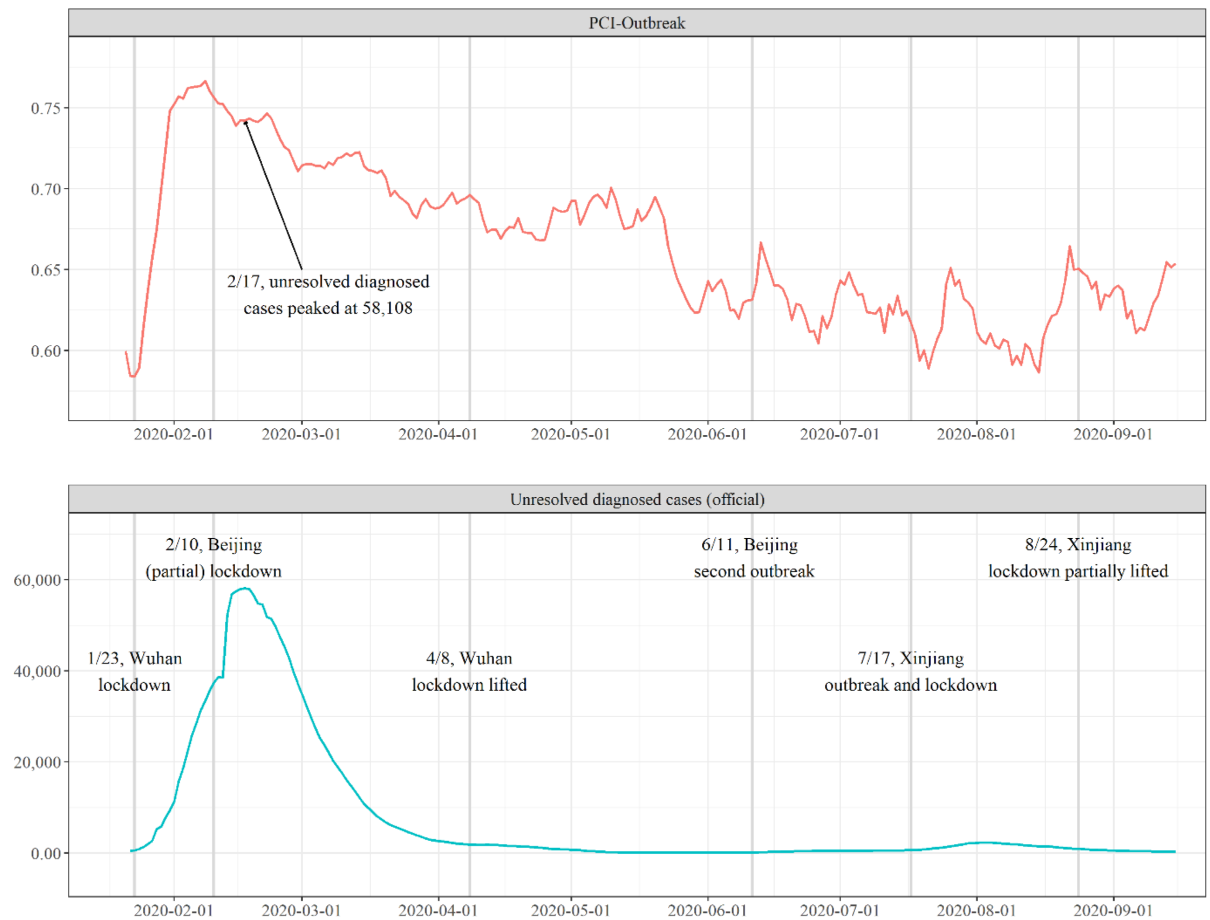
True week	Predicted week													
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.69	0.35	0.12	0.08	0.14	0.00	0.14	0.04	0.02	0.02	0.00	0.00	0.00	0.00
1	0.31	0.76	0.59	0.22	0.20	0.29	0.06	0.04	0.04	0.00	0.00	0.00	0.02	0.00
2	0.18	0.41	1.37	1.00	0.47	0.57	0.31	0.31	0.12	0.10	0.00	0.02	0.00	0.00
3	0.06	0.16	1.02	3.27	1.90	1.12	1.00	0.65	0.41	0.20	0.10	0.06	0.02	0.00
4	0.02	0.04	0.39	0.96	1.90	1.65	1.37	0.49	0.25	0.12	0.06	0.04	0.02	0.00
5	0.02	0.08	0.25	0.80	2.08	3.98	3.06	2.14	0.41	0.39	0.14	0.08	0.02	0.00
6	0.00	0.14	0.10	0.65	1.20	3.43	4.64	2.98	1.18	0.43	0.20	0.16	0.14	0.00
7	0.00	0.06	0.22	0.29	0.61	1.45	2.47	3.43	1.18	0.49	0.18	0.10	0.14	0.00
8	0.02	0.04	0.06	0.33	0.53	1.43	1.82	2.72	2.72	0.74	0.16	0.14	0.10	0.00
9	0.00	0.06	0.10	0.06	0.29	0.47	1.10	1.18	1.72	2.43	0.20	0.08	0.14	0.00
10	0.00	0.02	0.08	0.10	0.22	0.39	0.67	0.71	0.78	1.00	1.06	0.25	0.02	0.00
11	0.00	0.00	0.04	0.06	0.20	0.39	0.53	0.53	0.57	0.25	0.63	0.80	0.33	0.00
12	0.00	0.02	0.04	0.08	0.22	0.29	0.47	0.55	0.45	0.25	0.25	0.33	1.27	0.14
13	0.00	0.00	0.02	0.04	0.08	0.16	0.29	0.18	0.14	0.04	0.10	0.18	0.22	0.22

Note: The numbers are percentage points and they sum to 100.

5. Results

The main results are shown in figure 3, which plots the PCI-Outbreak (seven-day moving average) for COVID-19 in China from January 21, 2020, the day the *People's Daily* started to address the public health crisis, to July 9, 2020, the time of this analysis. The figure also plots the official numbers of unresolved diagnosed cases and indicates a set of events relevant to China's policy responses to the outbreak.

Figure 3. The Policy Change Index for Outbreak (PCI-Outbreak) for COVID-19 and Official Statistics



Although the scales of the PCI-Outbreak and the official numbers are different, comparing the shapes of the two curves yields three observations about the discrepancy between the Chinese government's own words and the statistics the government releases.

First, the two curves peak around the same time in February 2020. In terms of when the epidemic hits a plateau, therefore, the official numbers are consistent with the tone and tenor of the state media.[15] Although the official numbers reach their maximum about two weeks after the PCI-Outbreak does, this is primarily an artifact of the one-time release of more than 15,000 reclassified cases on a single day—cases which should have been attributed to the preceding days.

On February 2, 2020, for example, the *People's Daily* ran an article that celebrates frontline healthcare workers, calling them "soldiers in white scrubs."[16] The article, published around the time that the PCI-Outbreak shows a peak, signals resounding urgency. It describes frontline care as "a race against the God of Death," in which healthcare workers across the country pour into Wuhan, the epicenter, to help fight the outbreak. Some reportedly work their eight-hour shifts without eating or drinking. The narrative resembles what the newspaper showed in May 2003, when the SARS outbreak was also near its peak.

Second, the shapes of the two curves diverge markedly after they peak. Although the PCI-Outbreak trends downward from March to June, it declines at a much slower rate than the official numbers. In contrast, the official numbers have become negligible since April. This pronounced discrepancy suggests that the Chinese government's official statistics may have

[15] Consistency between these two measures does not imply that the official numbers reflect the actual scale of the outbreak, of course. But if the two measures are inconsistent, it is more likely that the official numbers are misrepresenting the truth.
[16] *People's Daily*, "Salute to the Soldiers in White Scrubs," February 2, 2020.

exaggerated the speed at which the virus was being contained—at least faster than its own words did.

News reports outside the *People's Daily* appear consistent with evidence that the speed of virus containment was exaggerated. In late March, for example, it was reported that Chinese Premier Li Keqiang, in a meeting on COVID-19 with local officials, warned them to not "conceal or omit information in pursuit of zero case reports" (*Bloomberg News* 2020). Because the *People's Daily* is a central-government publication while the (national) number of cases is the aggregate of locally reported statistics, the discrepancy between the two curves may have been in part driven by underreporting at the local level.[17]

Third, the PCI-Outbreak has remained elevated since June; the indicator is consistently well above 0.6. Moreover, the indicator is shown to have responded rather sensitively when regional outbreaks were reported again in recent months, such as during the second wave of cases in Beijing in June and the sudden spread of cases in Xinjiang from July to August. In contrast, the official numbers record only minimal movements. The discrepancy between the two measures suggests that these recent outbreaks may have been more severe than what the official numbers seem to indicate.

For example, on June 30, 2020, Beijing was in the middle of a lockdown in response to its second wave of new cases. An article in the *People's Daily* again reveals a sharp discrepancy between its language and the official numbers. It reiterates at length the necessity of the lockdown order. The article further justifies the lockdown by arguing that the purpose of strictly enforcing it is to "maximally ensure residents' health and demonstrate the high standard of virus

[17] In theory, an alternative explanation is that the local-level numbers of cases are accurate, but the central government uses words in the *People's Daily* that are disproportionately more serious than they need to be. This explanation, however, would be at odds with the established fact that local governments in China tend to fabricate economic data more than the central government does (e.g., Lyu et al. 2018; Chen et al. 2019).

containment in the capital city,"[18] in marked contrast to the numbers, which indicate only about a dozen new cases per day in a city with a population of over 20 million.

6. Discussion

In this section, we compare the PCI-Outbreak with conventional sentiment analysis, and we show that the former is more informative about the COVID-19 outbreak's severity than the latter. We also demonstrate that our main results are robust by investigating variations of the PCI-Outbreak algorithm.

6.1. Sentiment Analysis

Because the PCI-Outbreak is constructed from the *People's Daily*'s language, one may wonder whether conventional sentiment analysis could achieve the same goal and how such analysis compares to our findings. For example, does the newspaper's sentiment worsen as diagnosed cases rise but turn better after the outbreak hits a plateau?

To consider this alternative approach, we compute the sentiments of all outbreak-related sentences in the *People's Daily* articles using Amazon Comprehend, a tool developed by Amazon Web Services. It takes a sentence, which can be in Chinese, and it outputs four sentiment measures: the extent to which the sentence is positive, negative, mixed, and neutral. There are 177,934 sentences from January 22 to September 15, 2020, in COVID-19-related articles, and the sentiment analysis yields several observations.

First, most sentences do not indicate a clear sentiment: 93.26 percent of them are deemed neutral and a further 0.04 percent are considered mixed. Only 4.62 percent of the sentences carry

[18] *People's Daily*, "Beijing to More Strictly Manage Quarantine and Surveillance," June 30, 2020.

positive sentiments, and the remaining 1.70 percent are negative.[19] In other words, using sentiment analysis to gauge the Chinese government's view on the outbreak means discarding most of the text to begin with.

Second, sentiment in the *People's Daily* is less pronounced when the outbreak becomes more severe. The Pearson correlation coefficients between official case numbers and the positive and negative sentiment scores are −0.036 and −0.004, respectively. Similarly, the coefficients between the PCI-Outbreak and the sentiments are −0.181 and −0.121, respectively. Therefore, when the outbreak gets worse—as indicated by the official numbers or by the PCI-Outbreak—the newspaper's language turns more "emotionless." This again suggests that sentiment is not an informative barometer.

Third, using the Granger test (Granger 1969), we find that sentiment scores are "Granger-caused" by both the PCI-Outbreak and the official statistics. Table 5 lists the results of the Granger test: models (i)–(iv) show that the PCI-Outbreak drives the sentiments, not the other way around, and models (v)–(viii) show that the official statistics drives the sentiments, not the opposite. In other words, to the extent that sentiment still contains some information about the severity of the outbreak, it tends to be "after the fact."

Table 5. Granger Causalities between Sentiments, the Policy Change Index for Outbreak (PCI-Outbreak), and Official Statistics by Lagging Days

	Lagging days			
Model	1	2	3	4
(i) positive sentiment score	5.235*	1.781	2.089	1.098
~ PCI-Outbreak	(0.024)	(0.172)	(0.104)	(0.360)
(ii) negative sentiment score	9.860*	3.149*	4.615*	3.418*

[19] The fact that there are more sentences with positive sentiment than with negative sentiment reflects the typical style of the Chinese official media. Even during a public health crisis, the newspaper tends to show brighter spots, urging unity and praising frontline workers instead of stressing the (negative) impact of the disease.

~ PCI-Outbreak	(0.002)	(0.046)	(0.004)	(0.011)
(iii) PCI-Outbreak ~ positive sentiment score	2.741 (0.100)	1.678 (0.190)	1.354 (0.260)	1.741 (0.144)
(iv) PCI-Outbreak ~ negative sentiment score	0.389 (0.534)	0.442 (0.644)	0.929 (0.428)	1.507 (0.204)
(v) positive sentiment score ~ official number of diagnosed cases	7.936* (0.006)	4.336* (0.015)	3.151* (0.027)	1.926 (0.110)
(vi) negative sentiment score ~ official number of diagnosed cases	8.661* (0.004)	2.695 (0.071)	1.588 (0.195)	1.191 (0.317)
(vii) official number of cases ~ positive sentiment score	2.728 (0.101)	0.516 (0.598)	0.867 (0.460)	1.229 (0.302)
(viii) official number of diagnosed cases ~ negative sentiment score	0.665 (0.416)	0.064 (0.938)	0.195 (0.899)	1.058 (0.380)

Note: The numbers shown are *F*-statistics and the probabilities Pr(>F). Analysis was performed using the R package "lmtest."
* Indicates cases that have rejection probability of less than 5% ($p < .05$), in which the null hypotheses of the Granger test are rejected.
Source: Achim Zeileis and Torsten Hothorn, "Diagnostic Checking in Regression Relationships," *R News* 2, no. 3 (2002): 7–10.

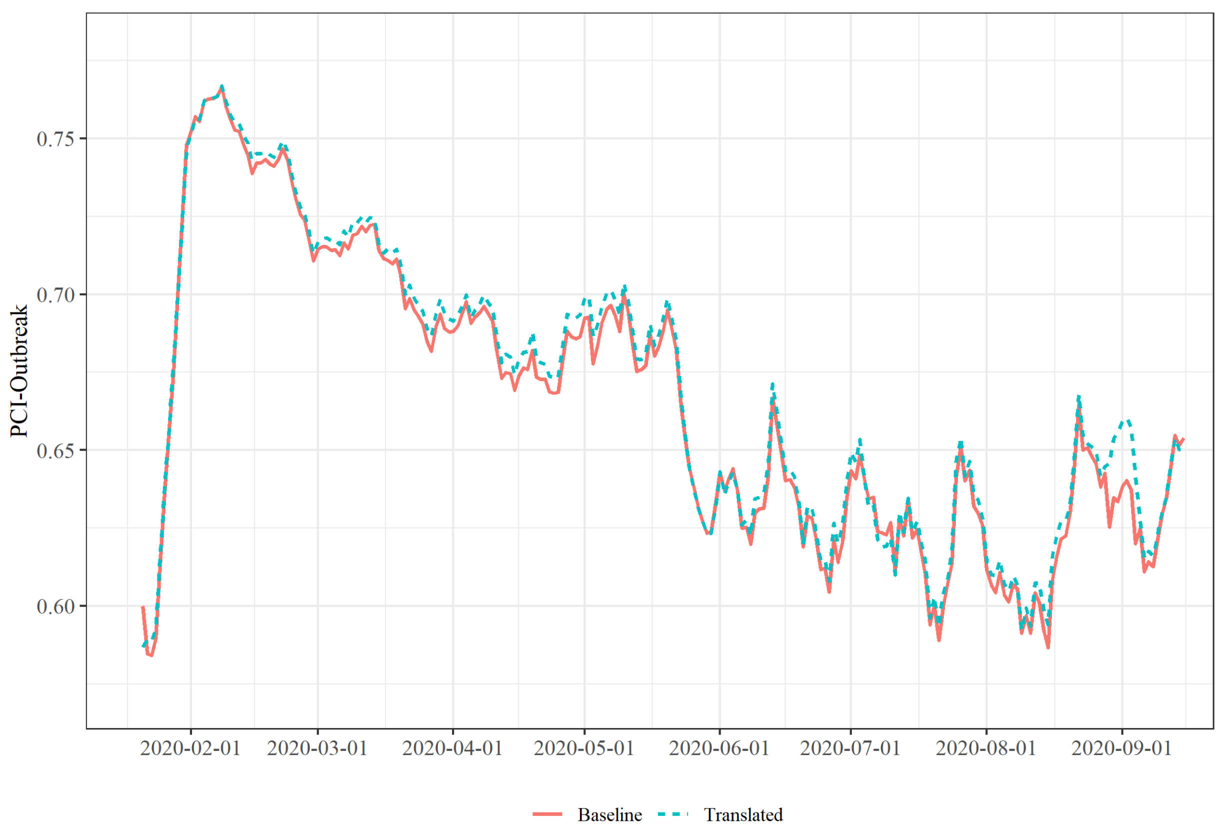
6.2. Robustness Checks

Because our analysis of COVID-19 relies on SARS as the benchmark but history does not exactly repeat itself, there are differences between the two episodes. Below, we discuss some of these differences and show that they are not central to the main results.

First, one may wonder whether the fact that the two episodes are driven by different infectious diseases matters. To answer this question, we have rerun the analysis by "translating" all COVID-19-related terms in the COVID-19-episode text to the respective SARS-related terms, so that the text is ostensibly about the SARS outbreak, even though it is not. For example, "pneumonia caused by a novel coronavirus," which was how COVID-19 was referred to by Chinese media at the beginning of the outbreak, is converted to "atypical pneumonia," its counterpart in the SARS episode. Similarly, "novel coronavirus" is converted to "atypical-pneumonia virus," "COVID virus" is converted to "SARS virus," and so forth.

Figure 4 compares the baseline PCI-Outbreak with the "translated" PCI-Outbreak. The difference between the two is negligible, suggesting that the nature of the disease does not matter to the severity measure.

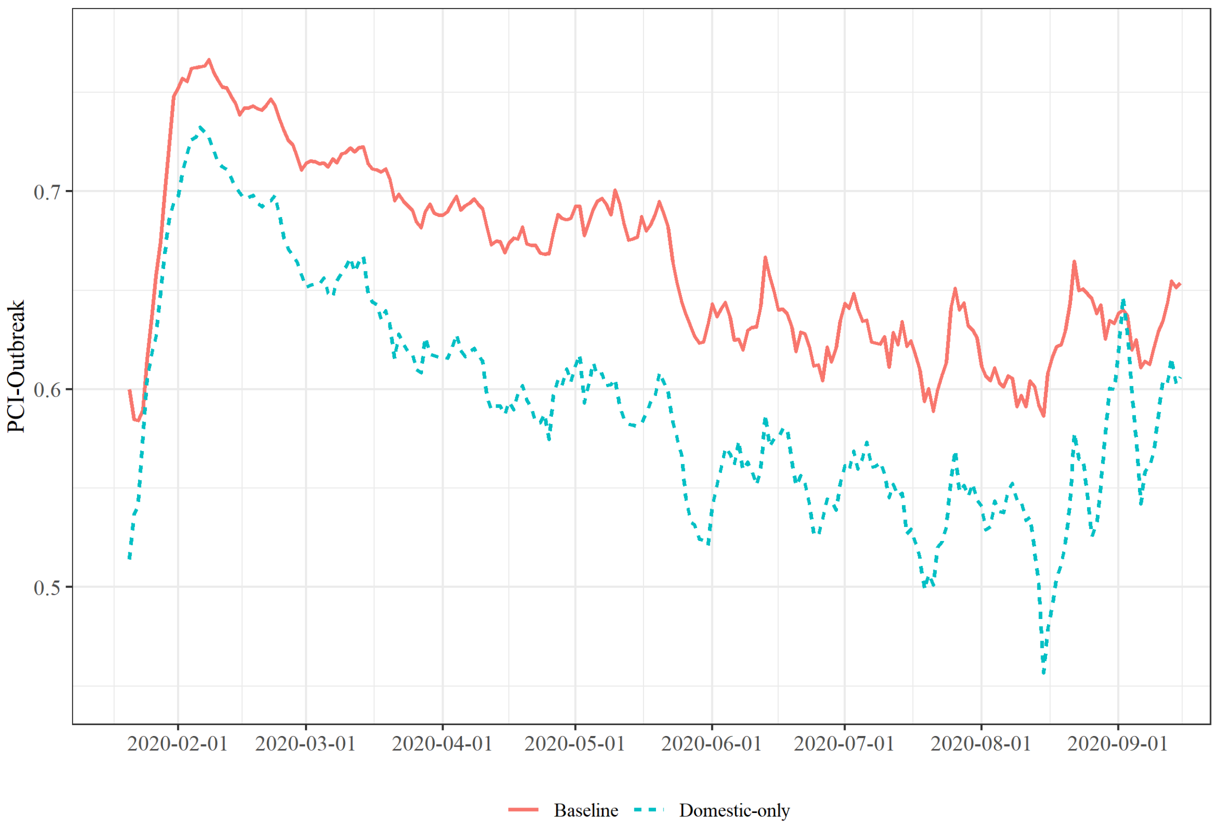
Figure 4. The Policy Change Index for Outbreak (PCI-Outbreak) for COVID-19, Baseline and Translated



Second, there is no question that the scale of COVID-19 is larger than that of SARS, because the latter did not lead to a global pandemic. Consequently, the *People's Daily* has published more coverage of outbreaks in other countries during the COVID-19 episode than during the SARS episode. It is possible, therefore, that the PCI-Outbreak remains elevated because, in the later stage of the COVID-19 outbreak, the severity picked up by the indicator is in reference to the virus's spread in other countries rather than in China.

To investigate whether this alternative explanation is valid, we have rerun the analysis by filtering out all overseas coverage of outbreaks, in both the SARS and the COVID-19 episodes. The difference between the baseline and the domestic-only indicators is shown in figure 5.

Figure 5. The Policy Change Index for Outbreak (PCI-Outbreak) for COVID-19, Baseline and Domestic-Only



This comparison invites two observations. First, our main results still hold using the domestic-only indicator. That is, as in the baseline, (1) the severity peaks in early February, (2) it trends downward from March to June—but only slowly, and (3) it responds sensitively to the resurgence of new cases in recent months—more so than the baseline.

Second, the domestic-only indicator is generally below the baseline indicator, by about 0.1 for the majority of the cycle. Because the baseline case contains data on both overseas and domestic outbreaks, this discrepancy implies that the foreign coverage in the *People's Daily* is more negative than the home coverage. This finding is consistent with the widely held impression that the Chinese government has downplayed the severity at home but overemphasized the virus's spread overseas.

7. Conclusion

We have developed a measure of COVID-19 severity in China using deep learning methods and relying on the text of China's official newspaper, the *People's Daily*. We have shown that this tool detects a marked discrepancy in severity between our measure and China's official statistics. In particular, according to our measure, the COVID-19 severity in China declined from March to June much more gradually than the official numbers suggest. We have also shown that our method is more informative than a conventional sentiment analysis of the text.

We recognize that our algorithm is benchmark-specific; more precisely, it is specific to using China's SARS outbreak as the benchmark. Therefore, it would not be meaningful to apply the model to, for example, news coverage on COVID-19 outbreaks in countries other than China. However, the method of leveraging well-known historical examples to learn how the tone and tenor of certain text evolve obviously has wide applications. While some of us have conducted a previous study along these lines (Chan and Zhong 2020), we leave these interesting possibilities to future research.

References

Agrawal, Sharan, Elliott Ash, Daniel Chen, Singh Gill, Amanpreet Singh, and Karthik Venkatesan. 2017. "Affirm or Reverse? Using Machine Learning to Help Judges Write Opinions." NBER Working Paper, National Bureau of Economic Research, Cambridge, MA.

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective." *PeerJ Computer Science* 2: e93.

Bloomberg News. 2020. "China to Reveal Data on Coronavirus Infected with No Symptoms." March 30, 2020.

Brady, Anne-Marie. 2008. *Marketing Dictatorship: Propaganda and Thought Work in Contemporary China*. Lanham, MD: Rowman & Littlefield.

Chan, Julian TszKin, and Weifeng Zhong. 2019. "Reading China: Predicting Policy Change with Machine Learning." AEI Economics Working Paper 2018-11, American Enterprise Institute, Washington, DC.

———. 2020. "Predicting Authoritarian Crackdowns: A Machine Learning Approach." Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA.

Chen, Wei, Xilu Chen, Chang Tai Hsieh, and Zheng Song. 2019. "A Forensic Examination of China's National Accounts." *Brookings Papers on Economic Activity* 2019 (Spring): 77–141.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. "What Does BERT Look At? An Analysis of BERT's Attention." In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, edited by Tal Linzen et al., 276–86. Florence, Italy: Association for Computational Linguistics.

Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–86. Minneapolis: Association for Computational Linguistics.

The Economist. 2010. "Keqiang Ker-ching: How China's Next Prime Minister Keeps Tabs on Its Economy." December 9, 2010.

Esarey, Ashley W. 2006. *Caught between State and Society: The Commercial News Media in China*. New York: Columbia University.

George, Alexander L. 1959. *Propaganda Analysis: A Study of Inferences Made from Nazi Propaganda in World War II*. Evanston, IL: Row, Peterson.

Gerrish, Sean M., and David M. Blei. 2011. "Predicting Legislative Roll Calls from Text." In *Proceedings of the 28th International Conference on Machine Learning*, edited by Lise Getoor and Tobias Scheffer, 489–96. International Machine Learning Society, Bellevue.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (3): 424–38.

Guo, Fangjian, Charles Blundell, Hanna Wallach, and Katherine Heller. 2015. "The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, edited by Guy Lebanon and S. V. N. Vishwanathan, 315–23. San Diego: Society for Artificial Intelligence and Statistics.

Imai, Natsuko, Ilaria Dorigatti, Anne Cori, Steven Riley, and Neil M. Ferguson. 2020a. "Report 1: Estimating the Potential Total Number of Novel Coronavirus Cases in Wuhan City, China." Imperial College London.

———. 2020b. "Report 2: Estimating the Potential Total Number of Novel Coronavirus Cases in Wuhan City, China." Imperial College London.

Katz, Daniel Martin, Michael J. Bommarito, and Josh Blackman. 2017. "A General Approach for Predicting the Behavior of the Supreme Court of the United States." *PLOS ONE* 12 (4): 1–18.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. "Text Classification Algorithms: A Survey." *Information* 10 (4): 150–217.

Kraft, Peter E., Hirsh Jain, and Alexander M. Rush. 2016. "An Embedding Model for Predicting Roll-Call Votes." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by Jian Su, Kevin Duh, and Xavier Carreras, 2066–70. Stroudsburg, PA: Association for Computational Linguistics.

Krippendorff, Klaus, and Mary Angela Bock. 2009. *The Content Analysis Reader*. Thousand Oaks, CA: Sage.

Lasswell, Harold D., Nathan Lettes, and Associates. 1949. *Language of Politics: Studies in Quantitative Semantics*. New York: George W. Stewart.

Li, Ruiyun, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. 2020. "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-CoV-2)." *Science* 368 (6490): 489–93.

Lyu, Changjiang, Kemin Wang, Frank Zhang, and Xin Zhang. 2018. "GDP Management to Meet or Beat Growth Targets." *Journal of Accounting and Economics* 66 (1): 318–38.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *1st International Conference on Learning*

Representations—Workshop Track Proceedings, edited by Yoshua Bengio and Yann LeCun. Scottsdale, AZ.

Nay, John J. 2017. "Predicting and Understanding Law-Making with Machine Learning." *PLOS ONE* 12 (5): e0176999.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, edited by Alessandro Moschitti, Walter Daelemans, and Bo Pang, 1532–43. Stroudsburg, PA: Association for Computational Linguistics.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, edited by Marilyn Walker, Heng Ji, and Amanda Stent, 2227–37. Stroudsburg, PA: Association for Computational Linguistics.

Schonhardt-Bailey, Cheryl. 2013. *Deliberating American Monetary Policy: A Textual Analysis*. Cambridge, MA: MIT Press.

Scissors, Derek. 2020. "Estimating the True Number of China's COVID Cases." American Enterprise Institute, Washington, DC.

Sim, Yanchuan, Bryan R. Routledge, and Noah A. Smith. 2016. "Friends with Motives: Using Text to Infer Influence on Scotus." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by Jian Su, Kevin Duh, and Xavier Carreras, 1724–33. Stroudsburg, PA: Association for Computational Linguistics.

Stockmann, Daniela. 2013. *Media Commercialization and Authoritarian Rule in China*. New York: Cambridge University Press.

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. "CommonSenseQA: A Question Answering Challenge Targeting Commonsense Knowledge." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4149–58. Stroudsburg, PA: Association for Computational Linguistics.

Tsang, Tim K., Peng Wu, Yun Lin, Eric H. Y. Lau, Gabriel M. Leung, and Benjamin J. Cowling. 2020. "Effect of Changing Case Definitions for COVID-19 on the Epidemic Curve and Transmission Parameters in Mainland China: A Modelling Study." *Lancet Public Health* 5 (5): e289–96.

Waltl, Bernhard, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. "Predicting the Outcome of Appeal Decisions in Germany's Tax Law." In *Electronic Participation: 9th IFIP WG 8.5 International Conference, ePart 2017, St.*

Petersburg, Russia, September 4-7, 2017, Proceedings, edited by Peter Parycek, Yannis Charalabidis, Andrei V Chugunov, Panos Panagiotopoulos, Theresa A Pardo, Øystein Sæbø, and Efthimios Tambouris, 89–99. Springer.

Wu, Guoguang. 1994. “Command Communication: The Politics of Editorial Formulation in the *People’s Daily*.” *China Quarterly* 137: 194–211.

Wu, Joseph T., Kathy Leung, and Gabriel M. Leung. 2020. “Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-NCoV Outbreak Originating in Wuhan, China: A Modelling Study.” *Lancet* 395 (10225): 689–97.

Xu, Hu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. “BERT Post-training for Review Reading Comprehension and Aspect-Based Sentiment Analysis.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 2324–35. Minneapolis: Association for Computational Linguistics.

Yano, Tae, Noah A. Smith, and John D. Wilkerson. 2012. “Textual Predictors of Bill Survival in Congressional Committees.” In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Eric Fosler-Lussier, Ellen Riloff, and Srinivas Bangalore, 793–802. Montréal: Association for Computational Linguistics.

Zeileis, Achim, and Torsten Hothorn. 2002. “Diagnostic Checking in Regression Relationships.” *R News* 2 (3): 7–10.

Zhao, Shi, Salihu S. Musa, Qianying Lin, Jinjun Ran, Guangpu Yang, Weiming Wang, Yijun Lou, et al. 2020. “Estimating the Unreported Number of Novel Coronavirus (2019-NCoV) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early Outbreak.” *Journal of Clinical Medicine* 9 (2): 388-93.

Zirn, Cäcilia, Robert Meusel, and Heiner Stuckenschmidt. 2015. “Lost in Discussion? Tracking Opinion Groups in Complex Political Discussions by the Example of the FOMC Meeting Transcriptions.” In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, edited by Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, 747–53. Hissar, Bulgaria: INCOMA Ltd.