Title: GDELT dataset appendix for “Chinese state media Facebook ads are linked to changes in news coverage of China worldwide”
Authors: Arjun Tambe (1), Toni Friedman (2)
Date: January 14th, 2022
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix B: GDELT dataset

In this Appendix, we list details of the GDELT dataset, including search terms used to collect the tone and content of news articles. The GDELT dataset aggregates over 60,000 news sources from around the world (GDELT Project, 2018). Online news sources are attributed to a country using several pieces of information, including their websites' domain ending, the countries their reporting focuses on, the country in which their companies are incorporated, and the country in which their websites' domain names are registered (GDELT Project, 2018).

Data collection and processing

We used the GDELT Full-Text Search API to collect the data (GDELT Project, 2017). The “Timeline Tone” tool was used to collect data on the average tone of news articles in each country over time, while the “Raw Timeline Volume” tool was used to collect the number of articles matching the keywords in each country over time. The query extended from January 1, 2018 to December 31, 2020.

To capture a set of articles about China, we used the search terms “Chinese Communist Party,” “CCP,” “China,” “Chinese leaders,” “Chinese government,” “Chinese govt,” “Beijing,” “President Xi,” “Jinping,” and “Chinese minister.” This helped minimize the number of articles that mention China in passing. We also excluded from our dataset countries with a total population below 150,000 because they have sparse or unreliable GDELT data. We also excluded countries that are unavailable in the Ads Library API (Cuba, Syria, Sudan, North Korea, and Iran).

Topic areas were selected from themes mentioned heavily in the advertisements' content. The set of advertisements and their content is available online at https://doi.org/10.7910/DVN/KQ39K6. Keywords were selected in order to match the language used in Chinese state media's Facebook advertisements and to filter for coverage that was primarily positive towards China. In particular, keywords for COVID-19 were selected from a list of words, drawn from Molter & DiResta (2020), that distinguished the CCP's narrative on the pandemic from other media sources' narratives. For Xinjiang and Hong Kong, the terms “terror” and “riot,” respectively, were used heavily in Chinese state media reporting, and thus strongly suggested an article was favorable towards China. Keywords for Huawei's 5G network and China's domestic poverty alleviation efforts were also identified by examining words used disproportionately frequently in Chinese state media reporting. All queries also include the keywords used to identify coverage of China listed above.

Validation

We then validated our datasets using three separate measures. First, we had human coders annotate 243 articles from the GDELT dataset (using the same search terms and the “Article List” query type) as having either a positive or negative tone. Duplicate articles and articles whose source texts were unavailable were removed. The human coders provided binary labels instead of numerical scores, a common procedure used in this context (Hu & Liu, 2004; Khoo & Johnkhan, 2017). We compared GDELT's annotations against the human coders' annotations. Table 1 shows the confusion matrix for positive and negative tone ratings, along with the precision and recall within each category.

Table 1. *Confusion matrix with precision and recall scores for GDELT tone ratings. Green entries correspond to predictions for which the human and GDELT labels match.*

	Human: Positive	**Human: Negative**	**Precision**
GDELT: Positive	38.3%	11.1%	77.5%
GDELT: Negative	9.5%	41.2%	81.3%
Recall	80.2%	78.7%	

Precision and recall for both categories were above 75%. GDELT's total accuracy was 79.4%. These results were consistent with results in previous literature (Hu & Liu, 2004; Khoo & Johnkhan, 2017; Kundi et al., 2014) and are comparable to the intercoder reliability of human coders.

Second, we validated the topic-specific article volume measure. Our goal was to validate that the articles GDELT retrieved were, in fact, relevant to the topic areas we had identified (precision), and that GDELT reliably retrieved articles about the specific topic area (recall). We retrieved 337 articles (using the same search terms and the "Article List" query type) and had human coders annotate each one as either matching one of the 6 categories or as matching none of the categories. Figure 1 reports these results.

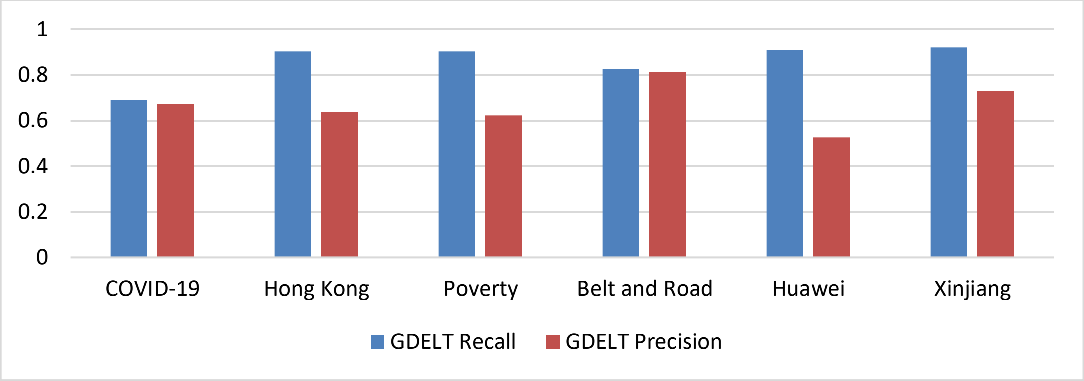

Figure 1. Recall and precision within each article category.

Recall was consistently high, above 80% for all topics but COVID-19. Precision was lower than recall, but remained above 60% for all topics except for Huawei. Precision may be lower than recall because the articles were typically classified as "other" by the human annotators if they listed several different issues, but were classified by GDELT as one of the topics if even *one* of the issues listed matched the topic-identifying keywords. When comparing the human coders' annotations against each other, the human evaluations agreed with each other 74% of the time; GDELT's performance was comparable to this figure. The high number of articles in our dataset also mitigated the impact of erroneous judgments (Shook et al., 2012).

Third, we used GDELT's tone scores to view the tone of the articles produced by each set of keywords. Figure 1 shows the average tone of coverage on China in general, compared to the average tone of coverage on each topic that matched the set of keywords. The error bars represent 95% confidence intervals (some error bars are too small to be visible on the graph). The graph shows that coverage that matched each keyword was much more positive, by at least 0.7 points, than coverage on China in general. This suggests that the keywords successfully filtered coverage on each topic to be more favorable towards China.

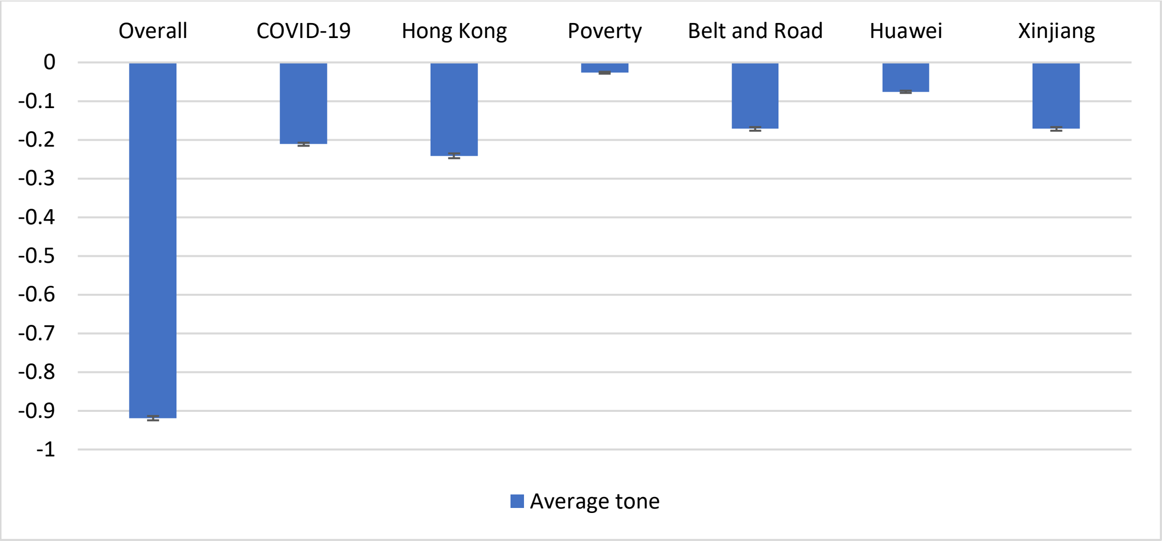

Figure 2. Average tone of articles that match keywords on each topic, compared to overall tone of coverage on China.