MERCATUS CENTER
George Mason University

RESEARCH SUMMARY

Words Speak Louder Than Numbers: Estimating China's COVID-19 Severity with Deep Learning

It is widely suspected that the Chinese government's official count of diagnosed COVID-19 cases understates the extent of the outbreak. In "Words Speak Louder Than Numbers: Estimating China's COVID-19 Severity with Deep Learning," Senior Research Fellow Weifeng Zhong and his collaborators develop a deep learning algorithm to estimate the outbreak's severity by analyzing the language of the *People's Daily*, China's official newspaper. They find a pronounced discrepancy between their severity measure, the Policy Change Index for Outbreak (PCI-Outbreak), and China's official number of cases.

PEOPLE'S DAILY AS AN INPUT, SARS AS A BENCHMARK

- The *People's Daily* serves as a suitable input for the algorithm because its text is tightly controlled by the Chinese government to promote official viewpoints and communicate top-down directives to the rest of the country.
- The authors use the 2002–2003 SARS (severe acute respiratory syndrome) outbreak in China as a benchmark and collect *People's Daily* text from both the SARS and COVID-19 outbreak episodes.
- The authors use a deep learning algorithm to learn where the tone and tenor of COVID-19-episode text would fit in the SARS-period articles' timeline. They then convert the fitted time of COVID-19-episode text to a measure of severity using the SARS record of diagnosed cases.

WHERE THE OFFICIAL NUMBERS AND THE PCI-OUTBREAK MEASURE DIFFER

- The Chinese government's official numbers of diagnosed cases are consistent with the PCI-Outbreak measure until the peak of the COVID-19 outbreak in China—both measures indicate that cases peaked in February 2020.
- After the peak, the official numbers and the PCI-Outbreak diverge. The PCI-Outbreak trends downward, but at a much slower rate than the official numbers, suggesting that the Chinese government may have exaggerated the speed at which the virus was contained.
- The PCI-Outbreak stays elevated from June to September, while the official numbers have stayed low since April. This discrepancy suggests that recent outbreaks in Beijing and Xinjiang may have been more severe than what the official numbers indicate.

KEY TAKEAWAY

The true scale of the COVID-19 outbreak in China may never be known. However, the detected discrepancy between the authors' measure and the government's official case numbers is an indication of possible underreporting. The pandemic is thus a sobering reminder that (1) the time has passed when China's opacity had few implications for other countries and (2) it is more important than ever for researchers and policymakers to put Chinese reports through filters to better understand current events in China.