CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media

Han Zhang & Jennifer Pan
Department of Sociology, Princeton & Department of Communication, Stanford

1. Motivation
- Political collective action is a powerful tool to challenge authoritarian regimes.
- Collecting data on collective action events in authoritarian regimes is extremely hard: traditional mass media are subject to strong regulation.
- Large-scale protest dataset in authoritarian regimes is scant.

2. CASM: Collective Action Events from Social Media
- Goal: identify offline collective action from social media data.
- Why social media: arguably the best source when news reports and government statistics are not available.
- Why difficult:
  - Short texts;
  - Extremely rare;
  - Posts about protests and posts about grievances are very similar (newspapers will pre-screen grievances).

3. CASM: Overview
- Obtain posts with protest-related words.
- Apply a two-stage deep learning classifier, using image and textual data, to identify posts about collective action.
  - First stage identifies posts about grievances.
  - Second stage identifies posts about protests.
- Identify unique collective action events from collective action posts, using location and time information.

4. CASM-China
- Implement CASM in China: high social media penetration and harsh media regulation.
- Last available government statistic: 85,000 protest in 2005.
- Data source: Sina Weibo (similar to Twitter).
- Training Data: 60000 protests, hand-curated from Internet sources by two human-rights lawyers (the Wickedonna Collection: goo.gl/Btr2zG).

5. CASM: System
- Output: 197,734 unique collective action events.

6. Internal Validation

7. External Validation with Publicly Available Datasets
- CASM-China identifies many more events than existing datasets (Jan. 1, 2016 to Jun. 30, 2016).
- CASM-China covers high proportion of events from existing datasets.
- CASM-China misses events in regions with limited internet and Weibo penetration (e.g., Tibet).

8. Validating Censorship Bias
Censorship does not bias CASM substantially:
- Censorship focuses on bursty events, but censorship rate is not 100%.
- Most collective action events are not bursty → low likelihood of censorship.
- Test with pre-censored data: 0.54% collective action posts later censored.

9. Contributions
- Deep learning using image and text together.
- Extensive internal and external validation.
- Reveals benefits and limitations of using social media for event data.
- Largest collective action dataset in any authoritarian regime.

10. Ethical Considerations
- "Dual-use" dilemma: our method and data could be used by malicious third parties.
- Make only event-level data public.