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

Supplemental Appendix

Han Zhang* Jennifer Pan†

Abstract

Protest event analysis is an important method for the study of collective action and social movements, which typically draws on traditional media reports as the data source. We introduce Collective Action from Social Media (CASM)—a system that uses convolutional neural networks on image data and recurrent neural networks with long short-term memory on text data in a two-stage classifier to identify social media posts about offline collective action. We implement CASM on Chinese social media data and identify over one hundred thousand collective action events from 2010 to 2017 (CASM-China). We extensively evaluate the performance of CASM through cross-validation, out-of-sample validation, and comparisons with other protest datasets. We assess the impact of online censorship, and find that it does not substantially limit our identification of events. Compared to other protest datasets, CASM-China identifies relatively more rural, land-related protests, and relatively few collective action events related to ethnic and religious conflict.

*Ph.D. Candidate, Department of Sociology, 107 Wallace Hall, Princeton University, Princeton NJ 08544

†Assistant Professor, Department of Communication, Building 120, Room 110 450 Serra Mall, Stanford University, Stanford CA 94305-2050; jenpan.com, (650) 725-7326.

1 Collective Action Coding Rules

Research assistants are asked to code a post as describing a collective action event if both of the following are true based on the text and/or image of the post:

- If there is a specific date and time for a group activity (note: "group" is defined as three or more people).
- If the group action is described (e.g., we're protesting / marching / demonstrating / group petition because...; there's a clash between a group and police) as happening in the real world with a specific location (e.g., town, village, or street).

Research assistants are told not to code any of the following posts as related to collective action events:

- If mobilization is only happening online.
- If the event is organized by the government, party, or state.
- If it is a group legal action (e.g., we're going to file a lawsuit). We do not consider group legal action to be contentious.
- If the post is vaguely hinting at past collective action ("past" is defined as more than 1 month ago).
- If the post contains grievances but contains no sign of actual physical gathering.
- If collective action takes place in other countries, even if it is Chinese people protesting.

In addition, research assistants are told that documentation of police brutality by itself does not constitute collective action, and simply having the word "protect rights" (维权) is not sufficient to label a post as being about collective action.

2 Selecting Keywords K

The first step for applying CASM and often other machine-automated event detection systems involves constructing a keyword dictionary in order to select relevant documents that are relatively rare from a large corpus (King et al., 2017). Our dictionary K contains the 50 most frequently occurring words in the Wickedonna Dataset, and we use K to construct the set of posts T_K, which all contain at least one of the keywords in K. Our keywords are:

> Homeowners; protect rights; migrant workers; hard-earned money; block road; military police; wage arrears; protest; ask for owed wages; banner; strike; marches; forced demolition; law enforcement; violence; violently demolish; surrounded by a mob; block the road; demonstration; township government; county government; district government; government gates; in front of government door; evil; gangs; collude; force; pollution; petition; arrest;

owed wages; onlookers; uphold justice; wage arrears; environmental protection; repression; legal rights; appeal; law of the land; defraud; truth; children; forcefully take land; redress an injustice; lawyers; petitioners; be responsible for the people; maintain stability; sit-in; forced land taking

业主,维权,农民工,血汗钱,堵路,特警,拖欠,抗议,讨薪,横幅,罢工,游行,拆迁,执法,暴力,强拆,围堵,拦路,示威,镇政府,县政府,区政府,政府门口,政府门前,黑心,黑社会,勾结,强制,污染,信访,抓走,拖欠工资,围观,主持公道,欠薪,环保,镇压,权益,诉求,王法,诈骗,真相,孩子,强征,申冤,律师,访民,为民做主,维稳,静坐,征地

Note that some Chinese terms have the same English translations.

We evaluate how our choice of K impacts data collection and the output of CASM in a variety of ways. The choice of K influences the data collection process because T_K expands with the size of K. Although using a larger dictionary expands the coverage of protest posts, it comes at the cost of time and low specificity. Figure 1 shows that the most 50 frequent words from the Wickedonna Dataset cover more than 86% of posts in the Wickedonna Dataset. If we increase dictionary size to 100, it only leads to a 4% increase in coverage of posts. If we increase dictionary size to 250, 95% of posts in the Wickedonna Dataset will be covered.

However, doubling the dictionary size will almost double the time it takes to collect posts that contain these words. Furthermore, because the most frequent words (e.g., protest) are usually more likely to be about collective action than the less frequent words (e.g., air pollution), a larger K would lead to a set of posts T_K that has lower specificity, which makes it more difficult for classifiers to correctly identify collective action events. Altogether, our analysis of the training data shows that a doubling in the time of data collection and lower specificity would only result in a four percentage point increase in recovery of relevant posts.

We examine how the choice of K influences what posts are recovered from the Wickedonna Dataset by year and by region because lacking a keyword may not be random with respect to characteristics of events. Figure 2 shows that the 50 most frequently occurring words in the Wickedonna Dataset has a relatively lower (67%) coverage of the posts in 2013, but achieves good coverage of the posts from 2014 to 2016 (over 80%). This is because the Wickedonna Dataset is heavily skewed toward data in later years, especially in 2015 and 2016, while posts in 2013 only contribute to 7.1% of all posts in the Wickedonna Dataset. This skewed distribution makes the 50 most frequent words more likely to characterize patterns of later years. Figure 3 shows that relationship between K and the coverage of the protests in the Wickedonna Dataset is less varied by region. Setting $K = 50$ recovers more than 80% of posts in all provinces, including in ethnic minority regions such as Tibet (西藏), Xinjiang (新疆), and Ningxia (宁夏).

We also evaluate the robustness of CASM's output to size of K. To do that, we create a subset of T_K that includes the top n keywords in K, and see how the output of CASM-China is impacted by the increase of k. Figure 4 plots the relationship between n and the events identified by CASM. The result shows that by expanding the number of keywords from 10 to 20, the number of events identified increases. However, as the size of dictionary grows larger and larger, the marginal increase in the number of events identified declines. If we expand the number of keywords from 40 to 50, there is little change in the number

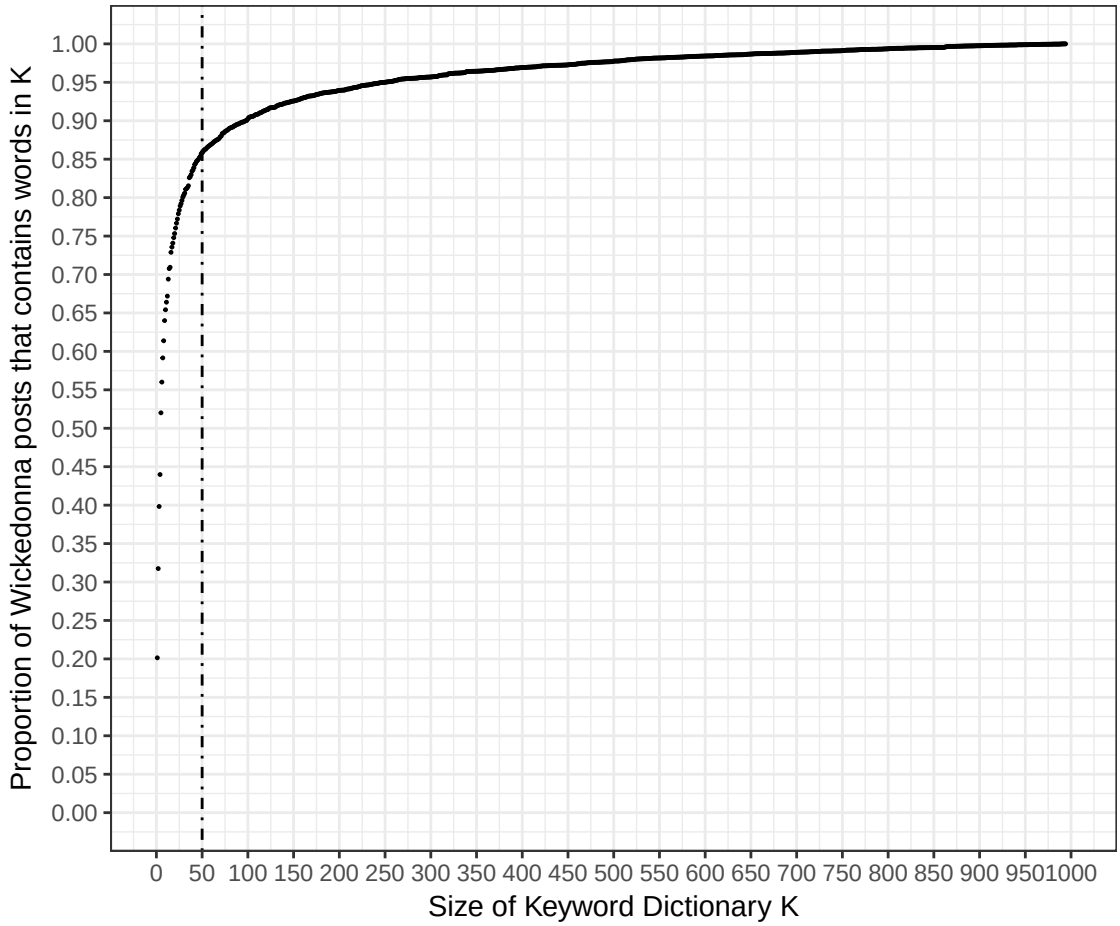

Figure 1: Coverage of protests in the Wickedonna Dataset by size of keyword dictionary; beyond 50 keywords, marginal coverage declines.

of events identified. The results suggest that by expanding the dictionary K beyond its current size of 50 is unlikely to substantially impact the identification of collective action events by CASM.

Finally, we show that expanding the size of dictionary leads to a decrease in our classifiers' performances. Figure 5, which shows the precision-recall curve for dictionary of size 10, 20, 30, 40, and 50 confirms this fact. CASM's performance is best when we only use the most frequent 10 words, and performance is slightly worse if the dictionary size is expanded to 50.

3 Model Comparison

Figure 6 shows how our CNN-RNN deep learning model (solid line) outperforms conventional supervised machine learning algorithms. We compare the CNN-RNN model from the first-stage with SVM and Naive Bayes. The input data for the SVM and Naive Bayes classifiers are identical to the deep learning models. We segment Chinese text. We only keep posts that have at least eight segmented words, and among these posts, we remove stopwords. In contrast to deep learning models, we determine how to represent the text

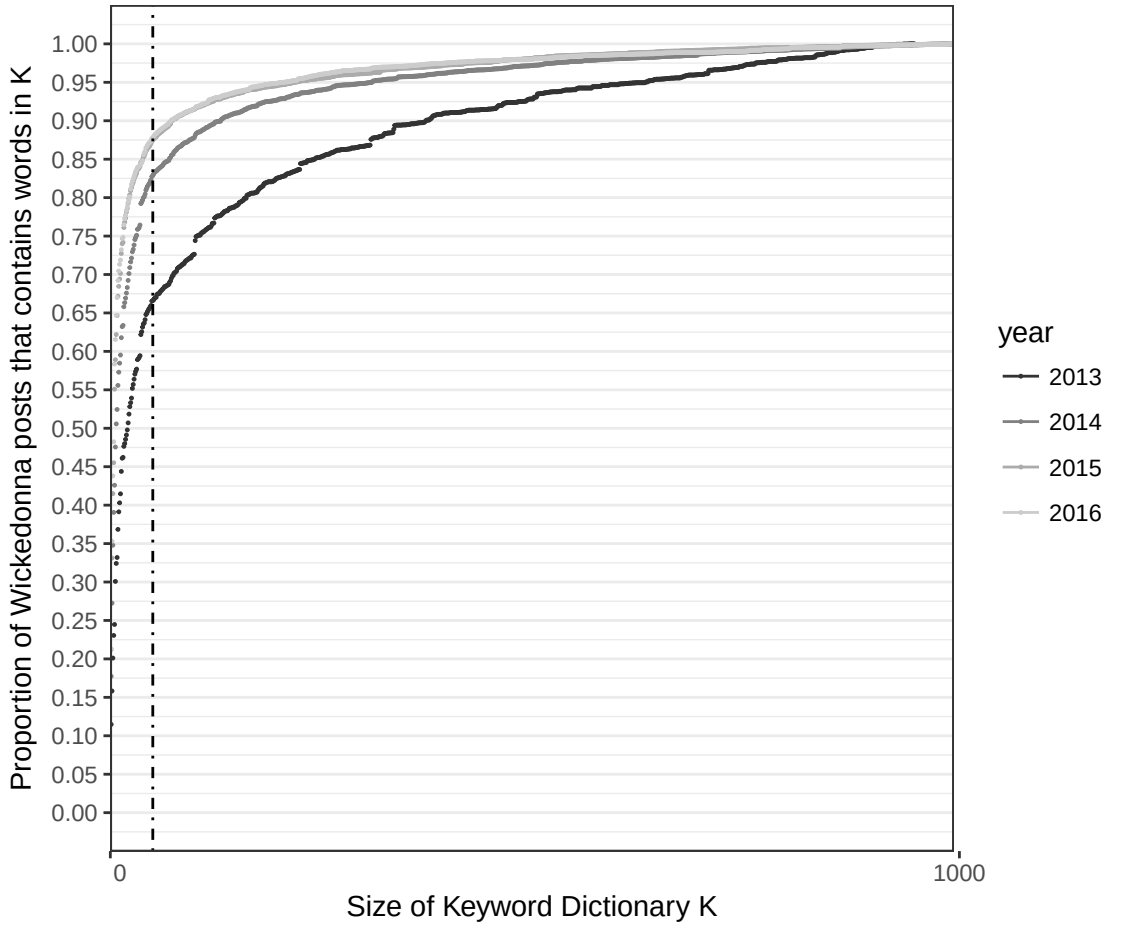

Figure 2: Coverage of protests in the Wickedonna Dataset by size of keyword dictionary, by year.

for SVM and Naive Bayes. We use n-grams with n ranging from $n = 1$ to $n = 5$, removing words that appear less than five times. We also use tf-idf transformation of the term-document matrix to down-weight frequent words while up-weighting rare ones.

4 Trends Over Time

The solid black line in Figure 7 shows the monthly count of events in CASM-China from January 1, 2010 to June 30, 2017. The number of events increases from 2010 to 2013, and slowly declines after. The 2013 to 2017 decline likely in part reflects the declining popularity of Sina Weibo. We see similar declines after 2013 in T_K, the number of posts containing the 50 collective action related keywords K (dashed line) and the number of posts containing a Chinese idiom we do not expect to relate to collective action also declines (dotted line).[1]

[1]The idiom is "half-hearted" (三心二意).

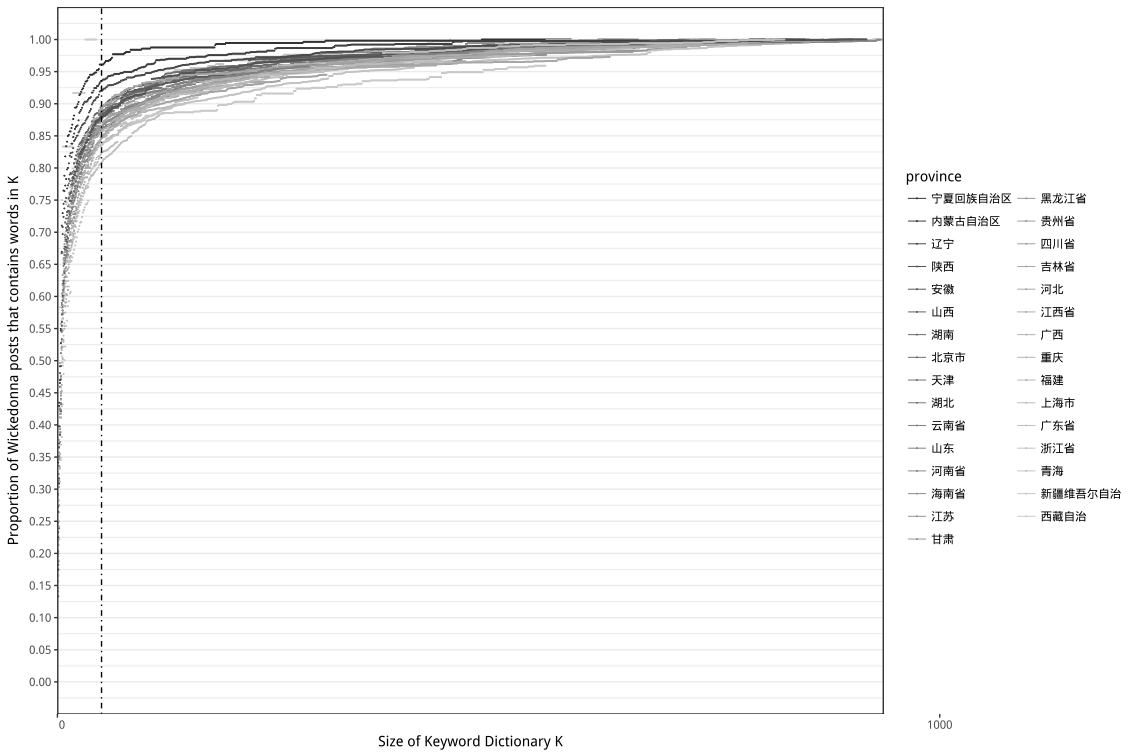

Figure 3: Coverage of protests in the Wickedonna Dataset by size of keyword dictionary, by provinces.

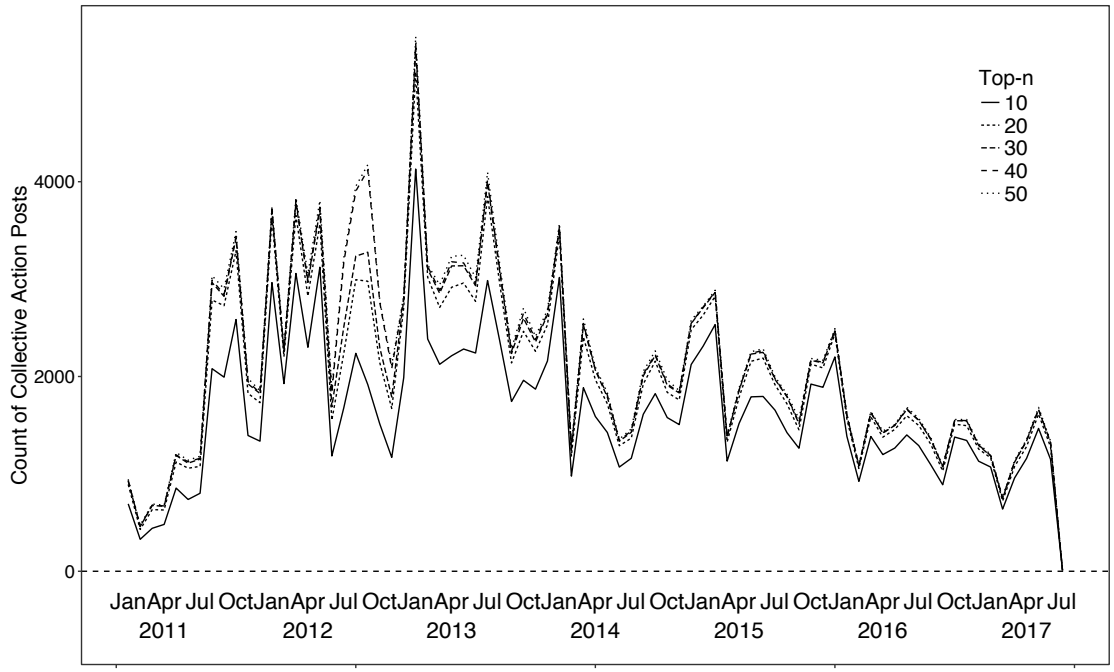

Figure 4: The number of collective action events identified in CASM-China by the size of the dictionary.

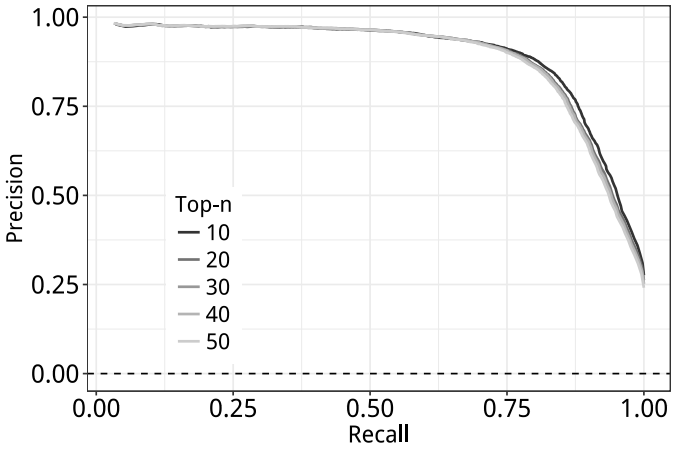

Figure 5: Precision-recall curve by size of keyword dictionary.

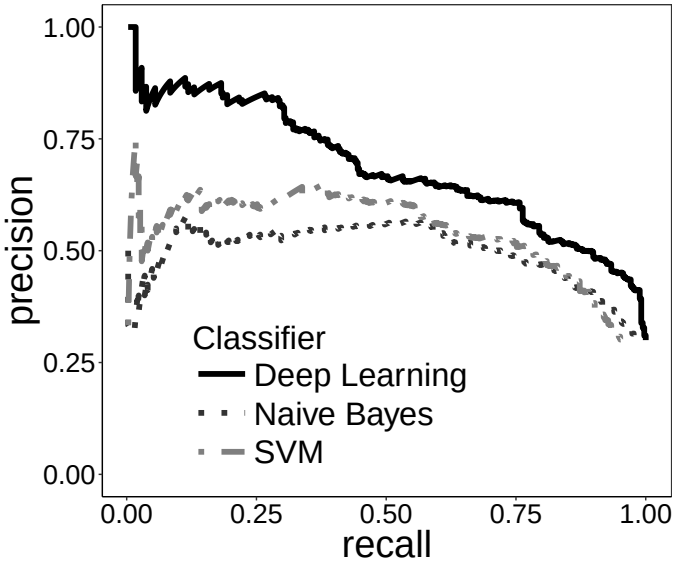

Figure 6: Precision-recall curve: deep learning vs. conventional algorithms. The comparison is based on text only.

5 Keywords to Identify the Form of Protest and Issues Motivating Protest

Our focus in this paper is to describe CASM and the main output of the system—the temporal-spatial distribution of protests. However, the text and images of $T_{protest}$ contain much more information about collective action events. Extracting this information in a rigorous manner is a priority for future research. We take a first-pass look at two features of collective action events—the form of protest and the issues motivating protest—using keywords. The keywords used to capture the form of protest are:

1. Conventional: “parade”, “strike”, “assembly”, “protest”, “voluntarily”, “upper level petition”, “name list”, “defend rights”, “petition”, “asking for back wages”, “arrears”, “arrears”, “pleading for help”, “hard-earned money”, “protect”, “interest”

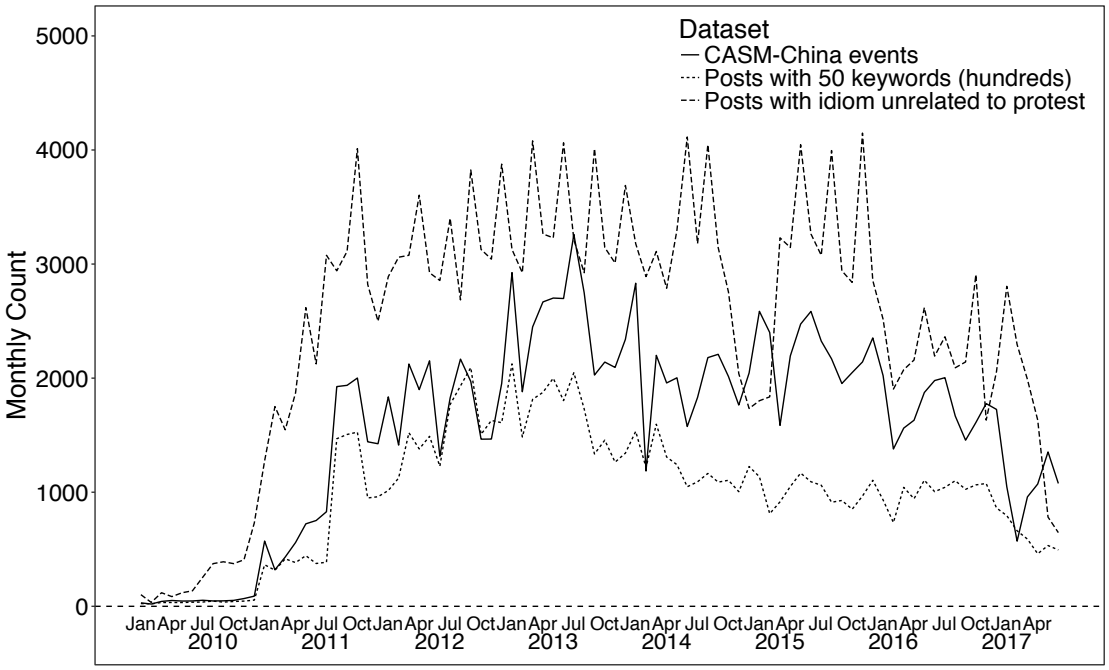

Figure 7: Monthly count of CASM-China collective action events (solid black line); monthly count of posts containing keywords K in hundreds (dashlined line); and monthly count of posts containing idiom unrelated to collective action (dotted line).

(“游行”, “罢工”, “集会”, “抗议”, “情愿”, “上访”, “签名”, “维权”, “信访”, “讨薪”, “欠薪”, “拖欠”, “求助”, “血汗钱”, “维护”, “利益’)

2. Disruptive: “block ”, “blocking”, “enclose”, “sit-in”, “doorway”, “blocking”, “banner”, “government front gate”, “disruption”, “make trouble”, “gather”, “block road”, “banner” (“堵”, “堵路”, “围堵”, “静坐”, “门口”, “封堵”, “横幅”, “政府门前”, “扰乱”,‘ ‘闹事”, “聚众”, “拦路”, “横幅”)

3. Violent: “attack”, “violence”, “armed police”, “police”, “beat down”, “police”, “forced demolition”, “forced”, “police vehicle”, “forced taking” “攻击”, “暴力”, “特警”, “警察”, “打倒”, “警务”, “强拆”, “强行”, “警车”, “强征”

The keywords used to capture the various issues (listed in alphabetical order) are:

1. Education: “parent”, “teacher”, “kindergarten”, “teacher”, “child”, “school”, “student”, “parent”, “education”, “go to school”, “high school”, “college”, “gaokao” [high-stakes college entrance exam], “college students” (“学生家长”, “老师”, “幼儿园”, “教师”, “孩子”, “学校”, “学生”, “家长”, “教育”, “上学”, “中学”, “大学”, “高考”, “大学生”)

2. Ethnic / religious: “Christianity”,“Catholicism”,“Protestantism, “church”, “Xinjiang”, “Uighur Minority”, “Uighur”, Islam”, “Muslim”, “Hui Muslims”, “Tibet”, “Tibetan”, “Tibetans”, “self-immolation”, “Islam” (“基督教”, “天主教”, “新教’,

”教堂”, “新疆”, “维族”, “维吾尔”, “绿教”, “穆斯林”, “回族”, “西藏”, “藏族”, “藏民”, “自焚”, “伊斯兰”)

3. Environment: “environmental protection”, “pollution”, “waste incineration”,“waste water”, “sewage”, “secondary pollution”, “chemical plant”, “refinery”, “air quality”, “burning coal”, “environment”, “soil environment”, “polluted”, “white pollution”, “smog”, “severely damage”, “nuclear radiation”, “sewage drainage”, “severe pollution” (“环保”, “污染”, “垃圾焚烧”, “废水”, “污水”, “二次污染”, “化工厂”, “精炼工厂”, “空气质量”, “燃煤”, “环境空气”, “土壤环境”, “污浊”, “白色污染”, “雾霾”, “严重破坏”, “核辐射”,‘ ‘排污”, “重度污染”)

4. Fraud / scams: “investors”, “scam”, “multi-level marketing”, “direct marketing, “fundraising”, “financing”, “defraud”, “profiteer”, “go bankrupt”, “people are gone buildings empty [idiom for scammers]”, “supplier”, “ funding ”,“collateral”, “commerce bureau”, “distribution” (“投资人”, “骗局”, “传销”, “直销”, “集资”, “融资”, “诈骗”, “奸商”, “倒闭”, “人去楼空”, “供应商”, “资金”, “抵押”, “工商局”, “物流”)

5. Homeowner / property: “real-estate developer”, “homeowner”, “property management”, “homeowners committee”, “residential committee”, “homeowner”, “household”, “homeowner”, “sales department”, “sales”, “resident entryway / building”, “rental housing”, “building management”, “smashing buildings under construction”, “apartment building”, “community”, “violate agreement”, “housing development”, “house management bureau”, “real estate”, “soy pulp [by product of tofu making, refers to poorly constructed building with quality issues]” (“开发商”, “业主”, “物业”, “业委会”, “居委会”, “房主”, “住户”, “屋主”, “售楼部”, “售楼”, “楼门”, “出租房”, “楼管”, “砸盘”, “公寓楼”, “社区”, “违约”, “楼盘”, “房管局”, “房产”, “豆腐渣”)

6. Medical: “hospital”, “family member”, “resuscitate”, “hospitalization”, “death”, “mediation”, “patient”, “patient”, “medical dispute”, “medical”, “surgery”, “emergency treatment”, “critical care”, “send”, “failed resuscitation”, “life”, “life in danger”, “medical staff”, “doctor” (“医院”, “家属”, “抢救”, “住院”, “死亡”, “调解”, “患者”, “病人”, “医闹”, “医疗”, “手术”, “急救”, “救治”, “送到”, “抢救无效”, “生命”, “生命危险”, “医务人员”, “医生”

7. Pension / welfare: “laid-off workers”, “retirement pension”, “five insurance and one fund” [retirement insurance], “old-age pension”, “social welfare” (“下岗工人”, “退休金”, “五险一金”, “养老”, “社保”)

8. Rural / land: “forced acquisition”, “land acquisition”, “demolition”, “land”, “bulldozer”, “acquisition”, “forced occupation”, “forced collection”, “relocation housing”, “forced relocation”, “destroy house”, “village tyrant” “强征”, “征地”, “拆迁”, “土地”, “推土机”, “征收”, “强占”, “强收”, “安置房”, “强迁”, “拆房”, “村霸”

9. Taxi: “driver”, “taxi”, “ride-sharing driver”, “public transportation”, “taxi”, “bus” “司机”, “出租”, “的哥”, “公交”, “的士”, “公交”

10. Unpaid wages: “unpaid wages”, “owed debt”, “to recover”, “owed money”, “to get back”, “to dock”, “blood and sweat money” [idiom for hard-earned money], “work fees”, “demand pay”, “to owe”, “in arrears”, “renege on debts”, “to owe debt”, “demand fairness”, “outstanding debt”, “owed wages”, “demand repayment”, “hard work”, “workers” (“欠薪”, “欠债”, “追讨”, “欠钱”, “讨回”, “克扣”, “血汗钱”, “工程款”, “讨薪”, “欠账”, “拖欠”,‘ ‘赖账”, “欠债”, “讨公道”, “赊账”, “欠工钱”, “讨债”, “辛辛苦苦”, “务工人员”)

11. Veterans: “veteran”, “discharged from military service”, “servicemen”, “People’s Liberation Army” (“老兵”, “退伍”, “军人”, “解放军”)

6 Generating Datasets for Comparison

In this section, we describe our procedures for collecting and creating collective action events datasets in China that are used to assess the external validity of CASM-China. For each dataset, we describes how we constructed or cleaned the data in order to compare it with CASM-China, including how we calculate its overlap with CASM-China.

GDELT: The Global Database of Events, Language, and Tone (GDELT) project takes a fully automated approach that relies on natural language processing to identify events of interest, including collective action events. GDELT tracks major news agencies around the world as the target source. We extracted all 10,620 events in GDELT between January to June 2016 that fell under the category of “Protest” and occurred in China.

We find that coding errors in GDELT are substantial, due mainly to its fully automated nature. We first clean obvious errors, including assignment of incorrect location of protests and duplicated events (multiple IDs associated with the same event). After this cleaning, only 2,214 unique event IDs exist. We next train a group of human coders to further code a random sample of 200 events from the 2,214 events to see whether the GDELT event represents a collective action event under our definition. We find that only 27 among the 200 fulfill our definition of collective action events. The remaining 163 tend to be newspaper articles published by Chinese newspapers about collective action taking place outside of China, irrelevant reports that contain protest-related words, or memorial articles that discuss the 1989 Tiananmen Square protests. This suggests that in expectation, GDELT only identifies about $\frac{27}{200} \times 2214 \approx 299$ collective action events between January and June 2016. We find that 15 of the 27 protests (55.6%) in GDELT are also in CASM.

ICEWS: The Integrated Conflict Early Warning System (ICEWS) is a DARPA program that combines political event datasets with an early warning system based on existing events.[2] Similar to GDELT, ICEWS also monitors global news agencies, but places more emphasis on the accuracy of identifying events rather than documenting as many events as possible (Ward et al., 2013). We first extract events between January to June 2016 that fall under the ICEWS category for protest, and then select events whose target and source countries are both China. This only returned 28 events, and 25 of them fit with

[2]https://dataverse.harvard.edu/dataverse/icews

our definition of collective action. Based on hand coding, we find that 18 out of 25 events (72%) events in ICEWS are also in CASM.[3]

WiseNews-China: WiseNews-China is built upon the WiseNews Database, which provides full-text articles from over 1500 major national and local newspapers from China, Hong Kong, and Taiwan.[4] Shao (2017) used the WiseSearch Database to identify 5,708 protest events from 1998 to 2014, based on keyword-filtering and human coding. His dataset is not available to the public, so we created a WiseNews-China dataset of collective action events by applying our two-stage classifier to the WiseNews Database. The only difference is that WiseNews-China uses newspaper articles from WiseNews as the data source. We use the 50 keywords in K to search for matching articles in WiseNews, and then run classifiers C_1 and C_2 sequentially to identify collective action events.

WiseNews returned 264,938 articles between January and June 2016 that contain at least one word in K. We are able to download 16,276 random articles.[5] Among them, our classifier identified only 106 articles related to collective action. Based on human coding, only 84 of the 106 articles are about protests, and from this, 17 unique events were identified by human coders. This suggests that in expectation, WiseNews contains $\frac{17}{16276} \times 264938 \approx 276$ events between January and June 2016.

Wickedonna: Here, we discuss how we calculate the overlap of the Wickedonna Dataset and our dataset. We first extracted 38,752 Wickedonna events that sourced from Sina Weibo (out of 67,502 total). For 19,615 out of the 38,752 events (48%), the exact post that Wickedonna used to identify the protest are also in CASM-China. For the rest of the unmatched events, we create a sample of 200 events in Wickedonna from January to June 2016, and let human coders check whether they are in our dataset. We find that for 33% of the 200 events, there are other posts in CASM-China that are describing the same event, which means CASM-China and Wickedonna are identifying the same collective action event, but based on different posts. In total, this suggest that in expectation, 65% (48% + 52% $\times$ 0.33) of the events in the Wickedonna are covered by CASM-China.

For 31% of the 200 events (15.8% = 52% $\times$ 0.31 of the total population), we find that they contain words that are not in our keyword dictionary so that CASM does not collect them.[6] 20% of the 200 events (10.4% = 52% $\times$ 0.20 of the total population) are no longer available on Weibo, either due to self-deletion or censorship. The remaining posts (8.3% = 52% $\times$ 0.16) are potentially not found by CASM-China due to Weibo's engineering restriction. Weibo bans searches for words including "protests" and "strikes," and for words that are very popular, such as "government," or "migrant workers" Weibo only returns at most 1000 posts per search. We maximize the number of posts by restricting the time periods, but some limitations remain.

China Labor Bulletin Strike Map: China Labor Bulletin is a Hong Kong-based NGO that aims to help labor workers bargain with employers and advocate for their rights. One

[3]There are five events in Tibet in ICEWS and only one of them is in GDELT.

[4]http://www.wisers.com/en/

[5]We cannot download more due to the website's restriction.

[6]Note that 15.8% is based on the sample of 200 events. Earlier we found that 14% of all Wickedonna events lacked one of the 50 keywords.

of their projects is to catalog labor protests in China. Their data comes from two sources. First, China Labor Bulletin has regularly searched for protest-related keywords on Chinese social media since 2010, and manually adds events into their dataset. This accounts for 46.7% of their entire dataset. In addition, China Labor Bulletin has incorporated all labor-related protests from Wickedonna[7] betwen June 2013 to June 2016, which accounts for 53.3% of their entire dataset. For the period between January to June 2016, 81% of events in the China Labor Bulletin are from the Wickedonna Dataset. We coded a random sample of China Labor Bulletin strikes (200 events) and find that 75% of their events are in CASM-China.

References

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61:971–988.

Shao, Dongke. 2017. "The Construction and Application of Mass Incidents Database in China." *China Public Administration* pp. 126–130.

Ward, M.D., Andreas Beger, J Cutler, M Dickenson, Cassy Dorff, and Benjamin Radford. 2013. "Comparing GDELT and ICEWS event data." *Analysis* 21:267–297.

[7] http://www.clb.org.hk/content/lu-yuyu-and-li-tingyu-activists-who-put-non-news-news