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

Han Zhang† Jennifer Pan‡

September 9, 2018

Abstract

Collective action is one of the most effective forms of political participation available to the public. The study of collective action has benefited greatly from protest event analysis that draws on data from traditional media reporting. However, in some settings, such as authoritarian regimes, where measures of collective action would be especially valuable, the government suppresses traditional media coverage of collective action. In this paper, we create CASM (Collective Action from Social Media)—a machine-assisted system that uses social media data to identify collective action events occurring in the real world by using deep learning, image as data, and two-stage classification. We discuss the advantages and limitations of this new data source, including the effects of social media censorship. We consider how validation can help make computer science methods more usable for social science research. We discuss the ethical implications of our system and the data, which we plan to make public. We apply our system to China, and demonstrate strong internal performance as well as external validity compared to existing newspaper-based and hand-curated event datasets. We identify 197,734 unique collective action events from 2010 to 2017, creating one of the largest datasets of collective action events in any authoritarian regime.

Keywords: Collective action, deep learning, social media, event data

*Our thanks to Isabella Cai, Weronika Cheng, Hong Fan, Yu Ji, Yuju Lin, Yingdan Lu, Tara Parekh, Yanchen Song, Feiya Suo, Debnil Sur, Jiajing Xu, Jing Wang, Yao Yang, Yunhui Ye, Shirui Peng, Kezhen Zhao, Yichao Cui, Elise Jiang, and many others for superb research assistance; Charles Chang for his work in data collection; Arturas Rozenas, Matt Salganik, Joshua Tucker, Yu Xie and Pamela Oliver for many helpful comments and suggestions; and to the Stanford Cyber Initiative and the Stanford Center for International Development Data for Development Initiative for research support.

†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 Introduction

Collective action—protest or other forms of collective social mobilization by groups outside the government—is one of the most effective ways the public can make their voices and demands heard. The study of collective action has benefited from protest event analysis that draws on data from traditional media reporting—newspaper articles, news wires, television programs—to quantitatively assess the occurrences and features of these events across geographic boundaries and over time (Hutter, 2014; Jenkins and Eckert, 1986; Jenkins and Perrow, 1977; Koopmans and Rucht, 2002; Kriesi, 1995; McAdam, 1982; Tarrow, 2005).

However, the limitations of using news reporting to identify and study collective action events have been well documented by social movement scholars (McCarthy et al., 1996; Oliver and Myers, 1999; Earl et al., 2004; Ortiz et al., 2005). In addition, the reliance on news reporting to generate protest event datasets has left gaps in knowledge in settings such as authoritarian regimes where collective action is crucial because opportunities for representation are limited but where information on these events is suppressed by the government and censored in the media precisely because collective action can threaten the survival of the regime (Egorov and Sonin, 2011; McMillan and Zoido, 2004; Qin et al., 2018; Stockmann, 2013).[1] In these settings, independent measures of collective action would be highly valuable for numerous scientific and public policy purposes, but government controls on media have made answering even basic factual questions about collective action events a challenge. For example, collective action is a major concern for China, but media reports of collective action events are strictly controlled (Lorentzen, 2014). Chinese academics, citing internal resources, have estimated that China experienced 180,000 mass incidents in 2010 (Wong, 2012). These counts are difficult to interpret since we know little about how the government defines mass incidents, how the counts were tabulated, and whether this definition has changed over time. Qualitative observations and geographically limited case studies are helpful for improving our understanding of the tactics and outcomes of specific protests (Cai, 2010; Chen, 2011; Deng and O'Brien, 2013; O'Brien and Li, 2006; O'Brien and Stern, 2007; Perry, 2002, 2008), but

they do not provide information on the overall contours and characteristics of collective action.

The adoption of digital technologies and the increasing digitization of our lives provide new opportunities for scholars to learn about collective action and to complement what we already know about collective action from traditional media reporting. Digital technologies—the internet, social media, mobile platforms—allow individuals to act as broadcasters and to disseminate information on a much larger scale ((Diamond, 2010; Earl and Kimport, 2011; Edmond, 2013; Ferdinand, 2000). Social media has become an important venue for protesters to speak out and to mobilize, and reflects participants' own accounts of collective action events, which allows us to capture how participants describe their motives for mobilization. Social media data are digitized and relatively accessible for large-scale collection. Even with online censorship and propaganda, the digital traces left by protesters, bystanders, and commentators provide us with new ways of identifying collective action events in authoritarian regimes (King et al., 2013, 2014, 2017).

In this paper, we create CASM (Collective Action from Social Media)—a machine-assisted system that uses social media data to identify collective action events occurring in the real world. CASM identifies collective action events from social media posts by using keywords related to collective action to extract potentially relevant posts, by applying a two-stage deep learning classifier, using image and textual data, to identify posts about collective action events occurring in the real world, and by merging posts discussing the same event to identify unique collective action events.

In creating CASM, we make three main contributions for social science research. First, we demonstrate the advantages and limitations of social media data as a target source for protest event analysis. As social media has become an important venue for protesters to speak out, social media data reflects participants own accounts of these events. Social media data allows us to uncover collective action events that would otherwise have been much more difficult to uncover. This is especially true in authoritarian regimes, and our approach remains robust even when we take into account censorship of social media by authoritarian governments. However, not everyone who engages in collective action has a

social media presence, so using social media as target data source will generate data that is biased toward certain geographies and populations. Second, CASM entails methodological innovations in the generation of protest event datasets by using deep learning, image as data, and two-stage classification. The use of deep learning algorithms—convolutional neural networks and recurrent neural networks—allows us to analyze images as data and to avoid bag-of-words assumptions typically associated with the analysis of textual data. In turn, this, along with a two-stage classifier, allows us to distinguish between social media posts that describe offline collective action events from posts that express grievances on similar topics but do not manifest as offline collective action events. Third, we show how internal and external validation can help make the application of computer science methods more practical and usable for social science research. In addition, we also discuss the ethical considerations of our system and the data it generates.

We implement CASM for China (CASM-China) using social media data from Sina Weibo. We generate one of the largest datasets of collective action events in any authoritarian regime, identifying 197,734 events from January 1, 2010 to June 30, 2017. CASM-China does extremely well in identifying posts, as assessed through cross-validation and out-of-sample validation, and also does well in identifying unique collective action events. We find that despite the fact online censorship in China focuses on suppressing discussions of collective action in social media, censorship does not have a large impact on the number of collective action events identified through CASM. This is because online censorship in China focuses on removing discussions of collective action that have captured the public's attention (are bursty), but most social media posts about collective action do not receive much attention (are not bursty), and censorship of bursts is incomplete. In assessing the external validity of CASM-China, we find that the system will miss collective action events taking places in minority regions, such as Tibet and Xinjiang, where social media penetration is lower and more stringent internet controls (e.g., internet blackouts) are in place.

Our goal is to detect collective action events in the real-world. CASM was not created just for the sake of developing a new method, but created to generate data that scholars

can use to advance our understanding of contentious politics in authoritarian regimes. To do this, we will make the dataset generated through CASM-China publicly available. It is important to keep in mind that we by no means claim to be able to identify all collective action events, or even all collective action events reported on social media. That said, our approach does allow us to identify numerous collective action events that might otherwise have escaped notice.

We proceed in four sections. Section 2 describes how traditional media has been used in protest event analysis, and the challenges associated with using data from traditional media reporting. In Section 3, we provide an overview of CASM, we describe the utility of social media as a target data source for identifying collective action events, we discuss how we use CASM to create a research dataset, and we consider the ethical implications of our system. Section 4 shows how we apply CASM to China, and presents the evaluation of the internal performance of the system along with an assessment of its external validity. Section 5 concludes.

2 Identifying Collective Action with Traditional Media

Protest event analysis has been an important method for social movement research for decades.[2] Protest event analysis allows researchers to systematically assess the occurrences and features of collective action events across geographic boundaries and over time.[3] Protest event analysis is a workhorse technique used in the development of several major research traditions in social movement research, including the political process theory, resource mobilization, new social movements in Europe, and comparative studies of global and transnational activism.

Protest event analysis requires the creation of datasets that document collective action events. Datasets are usually constructed in three steps: 1) retrieve relevant documents from target sources, 2) identify collective action events from these relevant documents, and 3) extract features of the identified collective action events (e.g., size of a protest and issue motivating the protest).

The main target source for the creation of collective action datasets has been traditional

media, and in particular, newspapers and newswire press releases. Newspapers provide a readily accessible source of data compared to other types of data such as government records. Well-known examples of newspaper-based collective action dataset include the US-focused Dynamics of Collective Action (DoCA) that used data from *The New York Times* from 1960 - 1995 (McAdam and Su, 2002), the PRODAT project that used Germany newspapers from 1950 - 2001 (Rucht et al., 1999), and the European Protest and Coercion Data that used newspapers in 30 European countries[4] (see Rucht et al. (1999); Earl et al. (2004); Hutter (2014) for more complete reviews of newspaper-based protest datasets).

A challenge to using traditional media as a target source for protest event analysis stems from systematic biases in media coverage of collective action events (McCarthy et al., 1996). Newspapers are more likely to report on larger protests, or protests that are more sensational in nature. Certain news outlets are more likely to report on some types of protest than other types. Altogether, research shows that selection bias in newspaper coverage of protests leads to bias in datasets constructed based on newspaper data (Earl et al., 2004). To ameliorate some of these biases, researchers use multiple newspapers as target sources (Azar et al., 1972; Oliver and Maney, 2000; Nam, 2006). They have also augmented newspaper-based datasets by using other forms of media and non-media content such as television transcripts, activists' websites, and Google search records (Gamson and Modigliani, 1989; Almeida and Lichbach, 2003; Earl and Kimport, 2008), and government archives (McCarthy et al., 1996).

Using traditional media as a target source to study collective action in authoritarian regimes faces additional challenges. Authoritarian regimes often impose strict controls on news reporting through state ownership of media outlets, and use repression and co-optation of of private outlets to shape content (McMillan and Zoido, 2004). They also conduct surveillance and impose physical controls on domestic and foreign journalists, often limiting their movements and sometimes using intimidation to censor reporting (Bourgault, 2015; Freedom House, 2017; Hem, 2014). As a consequence, many collective action events that happen in authoritarian regimes are not reported on in traditional media, either by local or foreign outlets.

Studies of collective action in authoritarian regimes have relied primarily on qualitative observations and geographically limited case studies. A few studies have used newspaper data to analyze collective action in authoritarian regimes but are typically focused on the characteristics of the events which are reported on rather than the contours of protests overall (Beissinger, 2007; Rasler, 1996; Chen, 2011). More often, protest event analysis in authoritarian regimes is only possible after the regime has fallen, when scholars get access to previously inaccessible archives and information sources.

3 CASM: Collective Action from Social Media

In this paper, we create a deep-learning approach to identify collective action events with text and image data from social media: CASM (Collective Action from Social Media). Drawing from McAdam et al. (2003), we define collective action as an episodic, collective event among makers of claims where targets are political and economic power-holders; where claims, if realized, affect the interests of at least one of the claimants, and where the action of claimants is a contentious event with public physical presence involving at least three participants (for details and coding rules see Appendix A).

CASM makes three main contributions for social science research. First, we introduce a new type of target data for protest event analysis—social media data, and we demonstrate the advantages and limitations of social media data as a target source. Second, we create a new approach for identifying collective action events that makes methodological innovations by using image as data, two-stage classification, and deep learning. Third, we show how internal and external validation can be used to help make the application of computer science methods to social science domains more practical and usable for social science research. We also discuss ethical consideration related to CASM, and how we minimize potential harm this system could generate.

3.1 Social Media as a Target Data Source for Collective Action

Given the challenge of using newspapers as source data for identifying collective action events in authoritarian regimes, the global adoption of social media provides an alternative

data source for identifying collective action. Using social media as target data has unique advantages but also important limitations.

Advantages: First, since anyone can broadcast on social media platforms, social media data gives researchers access to a wider array of collective action events, including those widely varying in scale, compared to what can be extracted from traditional media sources. This means that using social media as a data source allows us to uncover protest events which would otherwise have been difficult to detect. This advantage is especially crucial in authoritarian regimes where social media has become an important venue for protesters to speak out and to mobilize (González-Bailón et al., 2011; Smith, 2013; Trentham et al., 2015; Yang, 2003).

Second, social media reflects participants' own accounts of collective action events, and allows us to capture how participants describe their motives for mobilization. We gain a clearer understanding of the grievances, problems, and issues that mobilize. In contrast, traditional media is a mediated channel that describes collective action events in ways that can differ from those of participants (Gamson and Modigliani, 1989). While participant accounts may be exaggerated, social media data allow us understand how participants describe and frame their activities.

Researchers are increasingly turning to social media to study substantive topics in contentious politics and social movements, but researchers have not used social media as a data source for the creation of protest datasets. Most studies have focused on specific collective action events—for example, the Occupy Movement and the Gezi protests, and analyze how these events are discussed on social media and the role of networks in mobilization (González-Bailón et al., 2011; Budak and Watts, 2015; Barberá, 2015).

Some studies have attempted to forecast collective action events using social media data. While the procedure for forecasting is similar to the procedure for identifying past collective action events, forecasting studies typically emphasize prediction of collective action events that will be covered by news media, rather than uncovering previously unknown collective action events after the fact. To date, social media data, unlike newspaper data, has not been used by academic researchers to create post-hoc collective action event

datasets.[5]

Limitations: There are two main limitations to using social media data to identify collective action events. First, social media data is not representative. Second, social media is subject to censorship, especially in authoritarian regimes.

Social scientists have documented multiple biases of social media data, among them non-representativeness. Even though social media penetration is increasing (Rainie et al., 2012), users of social media platforms still constitute a non-random sample of the population (Mislove et al., 2011). In terms of identifying collective action, non-representativeness is an issue in that not everyone who engages in collective action has a social media presence, and as a result, not all collective action events appear on social media. This means that using social media as a target source will only uncover collective action in places and among populations that use social media (or the particular platform from which data is being collected). For example, we would identify few collective action events from social media data in countries like Iraq, Libya, or Turkmenistan because of low internet and social media penetration. We would also find relatively few tweets from Twitter to identify protests in China because Twitter is largely in accessible in China, and what tweets are found would come from individuals who can circumvent China's block of Twitter.

That said, non-representativeness should not prevent scholars from using social media altogether as a source of data, just as it has not prevented scholars from using traditional news sources to identify collective action events. Social media data is unlikely to reveal all incidences of collective action, but they may reveal events neglected by traditional media sources, which can help advance our understanding of collective action.

The second challenge related to using social media as target data is that social media is subject to censorship, especially in authoritarian regimes. King et al. (2013) and King et al. (2014) show that online censorship in China is aimed at removing discussions of collective action. Given that social media is censored, and in particular discussions of collective action on social media is censored, how can social media data be used to detect collective action events? The answer to this seeming contradiction lies in the recognition that censorship of collective action events described in King et al. (2013) and King

et al. (2014)—also known as content filtering—is post-hoc, focused on bursty online discussions, and incomplete. Content filtering—the deletion of content after it has already appeared online—is not based on keywords. Indeed, King et al. (2014) find that keywords used to flag posts for review before they are publicly posted do not predict what posts end up being censored. Instead, censorship is focused on bursts of discussion. Only when collective action events garner a great deal of discussion and attention on social media is the content censored. This means discussions of collective action on social media that do not attract out-sized attention will remain uncensored. Finally, Roberts (2018) shows that even with censorship of bursts of discussion of collective action, censorship of social media is usually incomplete. In other words, while most posts in a burst of online discussion about collective action will be censored, a few posts often escape censorship.

Given what we know about censorship, we expect to be able to identify collective action events from social media that do not garner enough online attention to attract the attention of censors. We expect this to represent the great majority of online discussions of collective action events because of the vast amount of information being shared on social media and the relatively few events in any domain that end up "going viral" or attracting broader attention. We also expect to identify collective action events that gather a great deal of online attention where it is challenging for censors to remove all related posts. We expect to have the greatest difficulty identifying collective action events through social media data for events that just generate enough online attention to be censored but not so much that censors have difficulty removing all traces. In sum, while censorship may reduce the total number of collective action events identified through social media, censorship is unlikely to invalidate the usefulness of using social media to identify collective action events.

3.2 Deep-Learning Approach with Text and Image as Data

Existing methods of building collective action datasets from traditional media data have used what we call fully human approaches, fully automated rule-based approaches, and machine-assisted approaches. Building collective action datasets from social media poses a challenge for each of these approaches.

The quantity of social media posts is vast but posts containing discussions of collective action events is extremely rare. Out of a random sample of 20,000 geo-coded posts from Sina Weibo, we identified one post discussing a real-world collective action event, which implies that 0.005% of social media posts in China discuss protest. The low proportion of collective action-related posts among the vast pool of social media posts makes the identification of collective action events that relies on fully human approaches impractical—to identify 100,000 protest-related posts, 2 billion social media posts would need to be coded, and assuming a person could code two posts a minute, this effort would entail 16 million hours of work.

The brevity and informality of social media posts poses a challenge for fully automated approaches, which for the most part have relied on the use of pre-define rules (based on either keywords, parts of speech tagging or pre-defined grammatical phrases) that machine apply to find matching content.[6] The use of slang and informal language on social media as well as the changing styles and uses of language on social media decreases the success of these fully automated rule-based methods Saraf and Ramakrishnan (2016).

Machine-assisted approaches refer to supervised machine learning, where human annotators create hand-coded training data, and then machine learning algorithms such as Support Vector Machine (SVM) or Naive Bayes learn patterns from these human-coded data. The trained algorithms then make predictions on whether new data discusses collective action events or not. Existing supervised-learning approaches are more adaptive to different data sources and more flexible than the rule-based approach (Nardulli et al., 2015; Hanna, 2017; Croicu and Weidmann, 2015), and they outperform rule-based methods in identifying collective action events based on newspaper articles (Ramakrishnan et al., 2014).

However, existing supervised-learning algorithms have difficulty identifying collective action events on social media. The key reason is that social media contains messages about collective action events that are occurring or have occurred, which use words similar to messages expressing grievances that could develop into collective action events but have not. In contrast, traditional media sources rarely report on failed mobilization, and

by extension on the grievances that could have or may evolve into collective action events but have not done so. Since the word usage of messages that discuss collective action events are often very similar to those that express grievances, conventional supervised learning methods often cannot pick up on the subtle differences needed to differentiate between these two types of messages.

CASM Overview: We also utilize a machine-assisted approach that uses supervised machine learning to identify collective action event. However, our supervised learning approach—CASM—differs from other, existing machine-assisted approaches because we tackle the methodological challenges associated with social media data by using image as data, two-stage classification, and deep learning.

Let T represent posts from social media. The goal of CASM is to identify collective action events, E, from T. CASM does so in three main steps:

1. Use keywords K related to collective action to extract potentially relevant posts T_K from T.

2. Apply a two-stage deep learning classifier, using image and textual data, to identify posts about collective action $T_{protest}$.

3. Identify unique collective action events E from $T_{protest}$.

Below we provide a description of each step with a focus on how the approach overcomes the challenges inherent to social media data.

Step One: We begin by constructing a dictionary of keywords K to identify posts from T that contain one or more of these keywords. We call this subset of posts containing keywords T_K (the outermost circle in Figure 1). Filtering the population of social media posts by keyword is essential because of the rarity of social media content related to collective action, but it is not sufficient because even among post containing protest-related keywords, few actually relate to real-world collective action events. The dictionary of keywords can be curated by experts by hand, or it can be calculated by identifying frequently

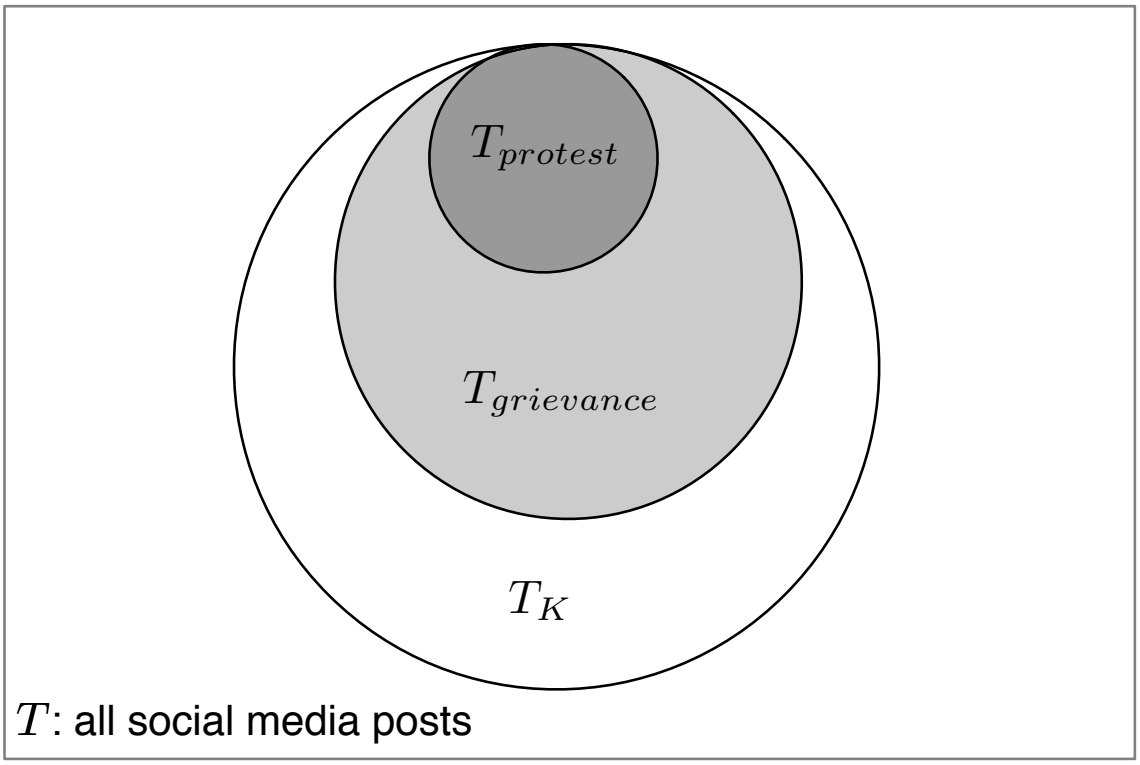


Figure 1: Social media posts discussing real-world collective action ($T_{protest}$) is a subset of social media posts that discuss ($T_{grievances}$), and $T_{grievances}$ is in turn a subset of posts containing keywords related to collective action (T_K).

occurring and/or differentiating keywords from social media posts known to discuss collective action.

Step Two: We then apply a two-stage deep learning classifier using text and image as data to identify posts from T_K, which discuss real-world collective action events ($T_{protest}$).

Image as data: Using text-as-data and conducting automated text analysis are increasingly common in social sciences (Grimmer and Stewart, 2013), and text-based content extraction has been the major workhorse in constructing protest event datasets (Hutter, 2014). Rarely has image been used as data in social science research, and images have not been used to construct protest event datasets. But images have long been recognized as important to the study of social movements, and they can be used to identify collective action events and useful characteristics, such as the size of protest, the target of protest, and the form of protest (Jacobs, 1967; McPhail and McCarthy, 2004). Furthermore, images

often play a role in catalyzing mobilization, as they convey strong emotional messages (Tufekci and Wilson, 2012; Kharroub and Bas, 2016).

Using image data in addition to text for our purpose is key as it helps distinguish the subtle difference between social media posts that describe a collective action event that has occurred and posts that express grievances but do not manifest offline as collective action. For example, a post with text that says "Come march with me" could be an individual trying to organize collective action, or it could be an on-going collective action event. In contrast, a post with text that says "Come march with me" and a picture of a person marching with hundreds of other protesters is unambiguously a real-world collective action event. Note that image data alone is also likely to be insufficient to identify collective action events as there are many types of group-based behavior that can be found in social media images—a holiday party, a concert—that are unrelated to the type of collective action we are interested in capturing.

Two-stage classification: Based on our close reading of social media posts, we find that posts about collective action events contain grievances, but not all social media posts about grievances relate to real-world collective action events. This is illustrated in Figure 1, where the $T_{grievances}$ is the subset of posts from T_K that discuss grievances, and $T_{protest}$ is a subset of posts from $T_{grievances}$ that are posts with grievances manifesting as real-world collective action. This creates challenges for machine learning algorithms because similar words can be used to describe grievances that do not manifest in real-world action and to describe collective action related to those same grievances. Machine learning algorithms that use text as data learn based on the words present in different documents, and the words used in "I went to protest with all 20 of my neighbors because pollution has been so bad" are similar to the words used in "Pollution is so bad that it makes me want to protest with my neighbors" even though the former refers to collective action that has occurred in the real world while the second does not. Therefore, an one-stage classifier will often fail to distinguish online complaints not associated with real-world collective action and online complaints about the same topics which do describe real-world events.

To solve this problem, we design a two-stage classifier. The first-stage classifier (C_1) identifies posts that contains grievances in general ($T_{grievance}$), and rules out posts that contain protest-related keywords but are unrelated to grievances. Then, we train a second-stage classifier (C_2) to differentiate between posts that express grievances but show no sign of real-world collective action and posts that express grievances that have manifested in the real world as collective action.

Deep learning: We use deep learning algorithms rather than conventional supervised machine learning algorithms in order to do two things: 1) analyze images as data, and 2) to capture long-range dependencies between words in the text data. On the first item, research in statistics and computer science have made significant advances in the analysis of image data because of deep learning (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). Drawing from this literature, we extract information from images posted to social media by using a Convolutional Neural Network (CNN) with 4 layers.

On the second item, conventional supervised machine learning algorithms such as Support Vector Machine (SVM) and Naive Bayes analyze text as term-document matrices that disregard word order and context with what is known as the "bag-of-words" assumption. Even when n-grams are used, these algorithms can only capture short-range dependencies between words (e.g., relationship within five words), and miss long-range dependencies. In contrast, deep-learning techniques have been shown to be able to capture long-range dependencies between words in texts (Kim, 2014). Our text-base classifier uses Convolution Neural Network and Recurrent Neural Network (CNN-RNN) with Long-Short-Term-Memory architecture. We find that these deep-learning algorithms outperform conventional machine learning techniques such as SVM or Naive Bayes. This is not surprising since deep learning allows us to use all the information available in short social media texts, whereas conventional supervised learning techniques will discard information such as word order.

Step Three: Finally, we identify unique collective action events E by adopting a rule-based approach that utilizes the temporal, spatial, and text information contained in the

posts $T_{protest}$. Specifically, we combine posts in $T_{protest}$ from the same geography posted on the same day.

3.3 Creating a Research Dataset

The goal of CASM is to create a dataset of collective action events that social scientists can use. Other studies have applied machine learning algorithms and methods to identify protest and collective action, but these studies have focused on demonstrating the advantages of these methods rather than creating an output that is of practical use for researchers. In order to apply computer science methods to social science, we extensively test and validate the performance of CASM. Existing work that uses machine learning for classification typically only evaluate performance with cross-validation. We go beyond this in three ways. First, we test the internal performance of our system in identifying posts with out-of-sample validation (in addition to cross-validation). Second, we test the internal performance of our system in identifying events (not just posts), and third, we assess the external validity of the resulting dataset.

Internal performance - identifying posts: We test the internal performance of our system for identifying posts along two dimensions: 1) whether the content of posts identified by CASM fits with our definition of collective action event (precision), and 2) to what extent CASM can retrieve posts about collective action events among posts containing protest-related keywords (recall). We cannot ascertain "true recall"—to what extent our classifier can retrieve the underlying pool of posts about collective action found on all of social media because the rarity of posts about collective action make the creation of a human-validated dataset based on all social media posts unfeasible. In CASM, we wish to balance precision and recall so that our systems identifies as many real-world collective action events as possible. Following common practice, we use the F_1 score ($F_1 = 2 * \frac{\text{precision + recall}}{\text{precision * recall}}$) as a measure of the overall performance of the system.

We conduct cross-validation and out-of-sample validation to estimate the precision and recall of our system. Cross-validation is the dominant approach for evaluating machine learning systems of event detection (Nardulli et al., 2015; Hanna, 2017). The train-

ing data is split into k equal subsets (we use k = 5). Each subset is used to calculate precision and recall with the rest used for training, and this process is repeated k times. The advantage of cross-validation is that class labels are already known for the training data, such that scholars can directly estimate precision and recall without additional effort.

Cross-validation shows the performances of classifiers on the training data, but the training data could differ from the data that researchers ultimately want to apply the classifier on. For example, the positive training data used for CASM-China encompasses a broader definition of protest than our definition of collective action and draws from a broader range of data sources. Therefore, precision and recall based on cross-validation can provide a rosier picture of the algorithm performance than is warranted. We conduct out-of-sample validation by taking a sample of the data the system will be applied on, letting research assistants label this data, and calculating precision and recall based on this independent test data. The critical advantage of out-of-sample validation is that it mimics the context where the classifier will be used, thus providing a more realistic evaluation of the system.

Internal performance - identifying events: To assess the internal performance of CASM in identifying events, we validate CASM by comparing human-labeled results with what is predicted by CASM along two dimensions: 1) event count accuracy and 2) within-event assignment accuracy. Event count accuracy refers to whether our event-detection approach can detect the same number of collective action events as our human research assistants. Within-event assignment accuracy refers to whether we incorrectly assign posts to any particular collective action event, and is calculated as the proportion of posts correctly assigned to any event. If every post that describes the same collective action event is correctly assigned, within-event assign accuracy equals 1. If none of the posts that describe the same collective action is assigned to the corresponding event, assigned accuracy is 0.

External validity: We assess the external validity of CASM by comparing it against existing datasets of collective action, and against other datasets we generate by using

traditional media as target sources of data. Even if internal performance is high, the resulting dataset may be of limited usefulness if the biases and limitations of the data are not known. Every dataset has limitations and biases, and assessing external validity allows researchers to conduct analyses and interpret results appropriately.

3.4 Ethical Considerations

We have oriented our work to striking a balance between maximizing academic transparency and ensuring ethical integrity. Before conducting this study, we secured approval from our University IRB. As a further protection for human subjects, we will make event-level information available but not individual-level posts or other information.

However, some ethical considerations merit further discussion. Our work has the potential to affect individuals who are not in our research. The methods and resulting datasets could be used by other actors—such as Chinese government, other authoritarian regimes, and activists—to identify, suppress or mobilize collective action. This means CASM faces the "dual-use dilemma" because CASM is created for research purposes but could be used by other actors in potentially harmful ways (Miller and Selgelid, 2007; Selgelid, 2013).

Our work is unlikely to create additional harm for those who engage in collective action in China. The Chinese government has an extensive apparatus to monitor its population, ranging from grassroots surveillance networks to CCTV to digital trace data, which includes monitoring social media (Crandall et al., 2013; Denyer, 2018; Pan, 2015; Qin et al., 2017). When it comes to collective action events, the Chinese regime collects detailed information, beyond what might be revealed on social media, about these activities. In other words, the Chinese regime has much more extensive information on collective action events than what they would gain by implementing our system.

However, other authoritarian regimes, with less extensive state surveillance apparatus, could try to implement our system to identify social unrest. Non-governmental actors, such as activists, could also use CASM to identify and try to mobilize individuals who have joined protest or lodged grievances against authoritarian governments, putting these individuals at greater risk. Following recommended practices from The National Science

Advisory Board for Biosecurity on dual-use research[7], we minimize this possibility by only providing an overview of our approach, instead of providing our codebase or detailed instructions for replicating our machine-assisted system.

4 CASM-China

In this section, we describe how we apply CASM to China to create one of the largest datasets of collective action in any authoritarian regime. We begin by describing the three main steps of CASM, and then move to describing our evaluation of internal performance and assessment of external validity.

4.1 Application of CASM to China

Our source for social media data in China is Sina Weibo (hereafter Weibo), China's biggest microblogging platform.[8] Weibo is functionally similar to Twitter, which is not easily accessible from China. Users can post message up to 140 characters. They can mention or talk to other users, use hashtags, follow other users, and repost.[9] Weibo is a open platform where users do not have to follow another user to read their posts. Research has shown that Weibo offers critical advantages for individuals to make their grievances, concerns, and reflections public, especially during time periods when traditional media outlets are silenced by the government (Sullivan, 2014; Huang and Sun, 2014; Lei, 2016).

Step One: We construct a dictionary of protest-related keywords K to extract a sub-set of posts from Weibo (T_K) by identifying the 50 most frequently occurring words from an existing dataset of social media discussions of protest in China—the so-called "Wickedonna dataset" created by activists Yuyu Lu and Tingyu Li. Between June 2013 to June 2016, Lu and Li identified 67,502 protests in China from Sina Weibo, Tencent Weibo, Qzone, and other online platforms. They published a daily list of protest events on their blog,[10] and each protest is associated with a number of related social media texts, images, and sometimes videos. In total, the Wickedonna dataset contains 240,521 text-based posts, and 233,288 images.[11] We chose 50 keywords to balance the trade-off between the

coverage of posts about collective action with the cost of data collection and the performance of our classifier. We conduct a number of tests that validate the selection of $K = 50$ as our dictionary size (see Appendix B for more detailed discussion).

We collected all Weibo posts published between January 1, 2010 and June 30, 2017 that contain at least one of the words in K—approximately 9.5 million posts.[12] For each post, we collect its text, images (if there are any), as well as available meta data such as the time of posting, the number of reposts, the latitude/longitude of the post, etc.

Step Two: Next, we apply our two-stage deep learning classifier to identify posts from T_K which discuss real-world collective action events. The first-stage combines a text classifier and an image classifier. For the text classifier, we use approximately 200,000 posts from the Wickedonna dataset as our positive training data. We use a random sample of approximately 200,000 posts from Weibo as the first negative training data—this random sample of posts from Weibo is extremely unlikely to contain discussion of protest or even social issues since the general prevalence of these posts is less than 0.01%. We use approximately 450,000 posts containing keywords from K that have a low likelihood of being about collective action as the second training dataset.[13] We use these training data along with the deep learning algorithm described above in the first stage classifier. For the image classifier, we use a random sample of 10,000 images from the Wickedonna dataset as our positive training data. We use a random sample of approximately 10,000 images from Weibo as the first negative training data. The final predicted probability is obtained by calculating the weighted average of the output probabilities of the text classifier and the image classifier; the weights are obtained by cross-validation.

For the second-stage classifier, we start by having a team of specially trained research assistants code nearly 40,000 posts that are identified by the first-stage classifier as having high probability of relating to collective action. Among these 40,000 posts, those coded as grievances that do not manifest as real-world collective action events are used as the negative training data, and posts coded as grievances that do manifest as real-world collective action events are used as the positive training data.

From this two-stage classifier, we identify a total of 283,427 posts out of 9.5 million

that are likely discussing collective action between January 1, 2010 to June 30, 2017.

Step Three: Finally, we identify unique collective action events from posts. Some events are discussed in only one post, while others are discussed in multiple posts. For each post, we extract its geographic location down to the county level by using geo-location of the posts when it is available in the meta-data, and geographic information from the text of the post when meta data is unavailable. After we have identified the location, we use meta-data on the time of posting to combine posts that come from the same location and were made on the same day into a unique collective action event. From the 283,427 collective action posts, we identify 197,734 unique collective action events. On average, each collective action event is discussed in 1.43 posts, which suggests that CASM is able to recover collective action events that receive limited overall attention on social media.[14]

4.2 Evaluating the Internal Performance: Identifying Posts

We first evaluate the performance of our two-stage classifier in identifying posts discussing real-world collective events. The dotted line in Figure 2 shows the precision-

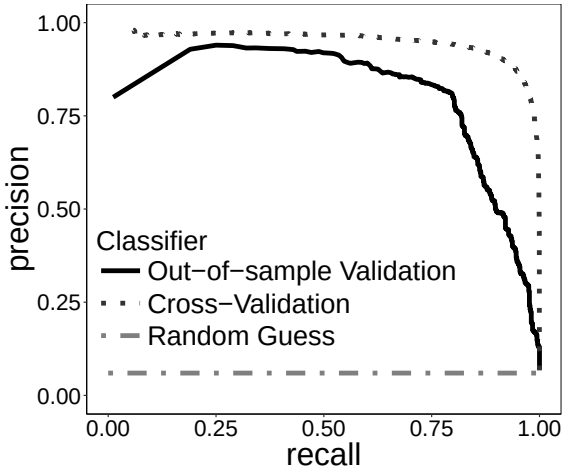


Figure 2: Precision-recall curve, out-of-sample validation vs cross-validation.

recall curve of the two-stage deep learning classifier based on image and text data from cross-validation. In cross-validation, CASM performs extremely well, with a maximum F_1 score of 0.94 (precision = 0.93, recall 0.94). This vastly out-performs random guess

based on the proportion of collective action show in the dotted-dash line in Figure 2. This also outperforms existing systems, which usually have F_1 scores between 0.6 - 0.8 (Hanna, 2017; Adams, 2014).

For out-of-sample validation, we take a stratified random sample of 200 posts per each keyword among from the 9.5 million posts collected between 2010 and 2017 (and these posts are not used during training). We have specifically-trained human coders to code each of the 10,000 sampled posts as discussing a action event or not per our definition. Then, we assess the performance of our classifier based on this independent validation set. As expected (Figure 2, solid line), performance is worse than in cross-validation, but still extremely strong with a maximum F_1 score of 0.83 (precision = 0.84 & recall = 0.80). This means that 84% of CASM's predicted collective action posts are correct, and CASM covers 80% of true collective action events from our out-of-sample validation data of 10,000 posts.

Our internal validation also reveals the importance of using image as data, of two-stage classification and deep learning to improve the performance of CASM. Figure 3

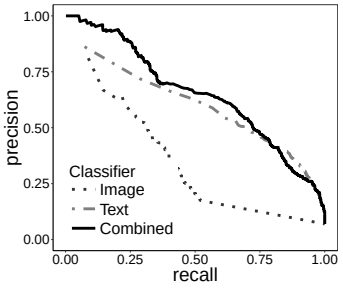


Figure 3: Precision-recall curve: text, image, and combined classifier (one-stage only).

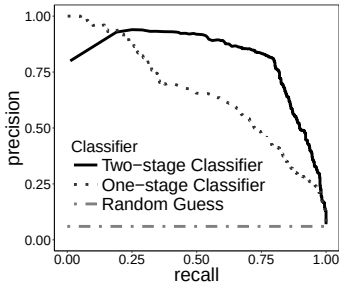


Figure 4: Precision-recall curve: two-stage classifier vs. one-stage classifier.

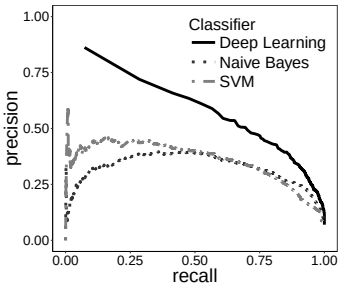


Figure 5: Precision-recall curve: deep learning vs. conventional algorithms. The comparison is based on texts only.[15]

shows that combining image and text classifiers (solid line) improves our ability to identify collective action events beyond using either text alone (dotted-dash line) or images alone (dotted line).[16] Figure 4 compares the two-stage classifier (solid line) with a one-stage classifier (dotted line). The two-stage classifier performs much better in terms of both precision and recall. Finally, Figure 5 shows how deep learning (solid line) outper-

forms conventional supervised machine learning algorithms, by comparing our one-stage classifier using deep-learning with SVM and Naive Bayes that use the exactly same pre-processed training data.[17]

4.3 Evaluating the Internal Performance: Identifying Events

To evaluate the performance of CASM-China in identifying events, we calculate event count accuracy by determining whether our event-detection approach can detect the same number of collective action events as our human research assistants. Our trained human coders identified 2,797 unique collective action events between January and June of 2016, while CASM identified 2,536 collective action events. Our approach identified 10% fewer collective action events than human coders because there can be multiple collective action events on the same date in the same county.

We also calculate within-event assignment accuracy, whether we incorrectly assign posts to any particular collective action event. We find that for 75.0% of collective action events in CASM, within-event accuracy achieves the perfect value of 1—every post that describes the same collective action event is correctly assigned by our rule-based approach as belonging to that event. For 18.9% of collective action events, assignment accuracy is 0—none of the posts that describe the same collective action are assigned to the corresponding event. This is because human coders identified more events than our rule-based approach. As a result, our approach assigned no posts to those unidentified collective action events. For the remaining 6.1% of collective action events, assignment accuracy is between 0.2 and 0.8.

Overall, these results show that our simple rule-based event detection approach works quite well in assigning posts to the corresponding collective action event.

4.4 Assessing External Validity

Now that we have established the strong internal performance of CASM-China, we assess the external validity of CASM-China. There are several factors that can limit the external validity and representativeness of collective action events identified through social media. First, government censorship of social media could mean that certain protests are missing.

Second, certain types of protest are not reported on social media. We would, for example, expect to uncover fewer collective action events in regions with lower levels of social media usage or lower adoption of Sina Weibo. Finally, an automated system is unlikely to recover all collective action events reported on social media because of the rarity of these events relative to the population of social media posts. We find that censorship is unlikely to have a large impact on the number of collective action events identified by CASM-China, but we find that CASM-China will miss collective action events in Western regions of China such as Xinjiang and Tibet, and CASM-China does not capture all collective action events reported on social media.

Censorship Discussions of collective action in China that garner online attention (bursty discussions) are censored (King et al., 2013, 2014). As discussed in Section 3.1, we expect censorship to have a limited impact on the ability of CASM to identify collective action events in China because most social media posts about collective action events do not receive much attention (are not bursty), and censorship of bursts is incomplete.

We empirically validate this expectation by examining censorship of posts related to collective action identified by CASM. We use a corpus of Weibo data collected in real-time, pre-censorship from January 2018.[18] Among these pre-censored posts, we find 113,081 posts containing at least one of the 50 keywords in K. We apply CASM to these 113,081 posts and identify 3,332 posts related to real-world collective action. Among these 3,332 posts, only 18 posts (0.54%) were later censored. This analysis shows that censorship is unlikely to have a large influence on the ability of our system to detect collective action events.

Comparison with Newspaper-based Datasets of Protest in China As discussed in Section 2, most protest event datasets are based on newspaper data. To compare our system based on social media data against systems based on newspaper data, we compare the output of CASM-China against two global newspaper datasets that cover protest in China (GDELT and ICEWS) and against one Chinese newspaper dataset, constructed by applying CASM to a large corpus of Chinese newspapers (WiseNews).[19] Details on these

three comparison datasets can be found in Appendix C.

Table 1 shows the comparison of CASM-China against these three newspaper-based datasets of protest between January and June of 2016. The first thing to note is that during this time period, the number of collective action events identified via CASM-China is hundred of orders of magnitude larger than the number of events reported in any newspaper-based data source. The low number of collective action events reported in both global and Chinese newspapers reflects the high level of control that China exerts over traditional media reporting.

Table 1: Comparison of CASM-China with Other Datasets of Collective Action in China (Jan. 1, 2016 to Jun. 30, 2016

	Source Data	Time Range	Number of Events Jan-Jun '16	Proportion of Events Covered by CASM Jan-Jun '16
CASM-China	Social media	2010-17	12,662	
GDELT	Int'l newspapers	1979-	299	56%
ICEWS	Int'l newspapers	1979-	25	52%
WiseNews	Chinese newspapers	1998-	276	88%
Wickedonna	Social media	2013-16	11,085	70%
China Labor Bulletin	Mixed	2011-	1,455	75%

Table 1 shows that GDELT identified 299 collective action events in China between Jan. to Jun. 2016, and ICEWS identified 25 between the time. Slightly over half of the collective action events identified by GDELT and ICEWS between January and June 2016 are also detected by CASM-China—56% of events in GDELT, 52% of events in ICEWS. The collective action events identified by GDELT and ICEWS but not found in CASM-China are all protests occurring in western minority-dominated regions of China, in particular in Tibet and Xinjiang. This result is not surprising since the Chinese government imposes more stringent internet controls in these regions, and usage of Chinese-language social media platforms is much lower in these regions than in the rest of China. This result highlights the importance of international reporting in China, especially in ethnic minority Tibetan and Ugyhur regions, but it also reveals the bias of international newspapers to focus reporting on protest in ethnic minority regions and to pay little attention to the large

numbers of protest occurring elsewhere in China.

Wisenews is one of the largest electronic, searchable databases of Chinese local newspapers, including both party newspapers and more commercialized newspapers (Shao, 2017). We apply our method for detecting collective action events to data from Wisenews to construct a newspaper-based dataset of collective action events in China. In comparison with data from the WiseNews corpus, CASM-China performs well, recovering 88% of 276 collective action events reported in local Chinese newspapers.[20]

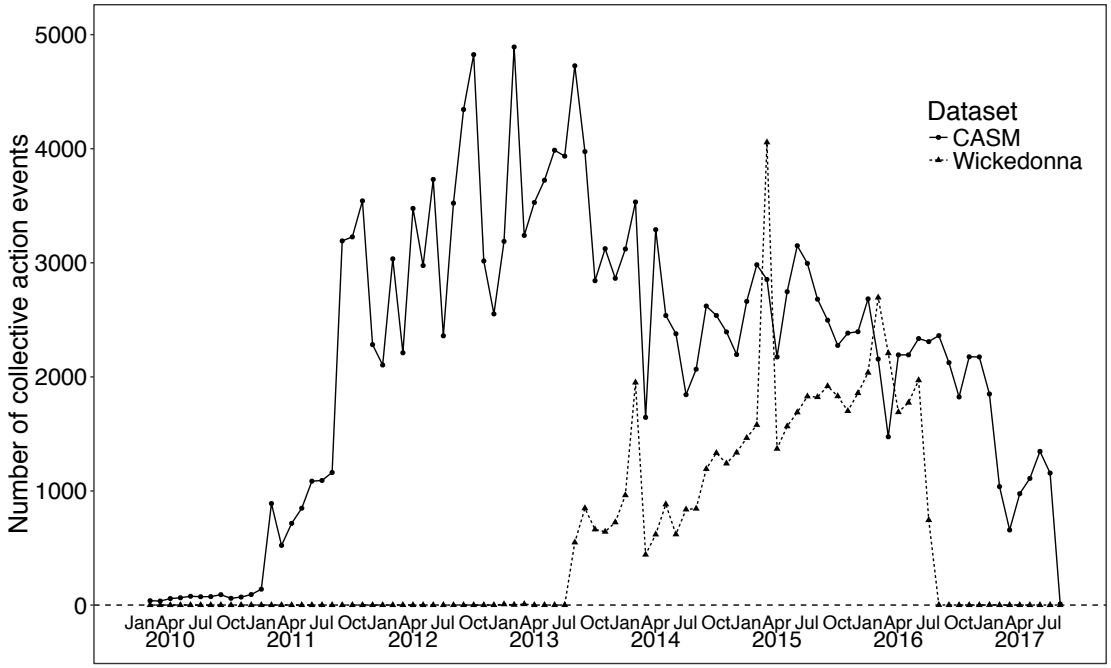
Comparison with Hand-Curated Datasets of Protest in China We compare CASM-China against two hand-curated datasets of protests in China: the Wickedonna Dataset, which is used as part of our training data, and the China Labor Bulletin, which documents labor-related protests.[21] Both datasets have been used by Chinese scholars to study collective action (Dimitrov and Zhang, 2017; Goebel, 2017).[22]

CASM-China covers many orders of magnitude more collective action events than the China Labor Bulletin dataset. Among protests reported in the China Labor Bulletin, 75% of events are covered by CASM-China. Since data for the China Labor Bulletin is based on the subset of data from Wickedonna, especially during the first half of 2016, and because Wickedonna is our training data and (prior to our work) the largest dataset of collective action events in China, we examine in greater depth how our data compares against the Wickedonna data.

Within the timeframe, 70% of events in the Wickedonna dataset are covered by CASM-China. Among the 30% of collective action events identified by Wickedonna that are not in CASM, 16.6% of the events in Wickedonna are not detected by CASM-China because they do not contain any keyword from our dictionary K; 7.3% of the events in Wickedonna not identified by CASM-China are posts no longer found on Sina Weibo, likely due to censorship. Finally, the remaining 6.1% are not found by CASM-China is due to Weibo's restriction on data collection.[23] These results corroborates our analysis of censorship that censorship is unlikely to have a large impact on event identification. These results also show that CASM-China by no means captures all collective action events reported on social media.

CASM-China only documents 1,577 more collective action events than the Wickedonna dataset between January and June 2016. This is because January to June 2016 is a time period where the Wickedonna dataset reaches its peak. Figure 6 shows that with few exceptions, CASM-China identifies significantly more events than the Wickedonna Dataset (especially during 2013 - 2014), and identifies events outside the scope of the Wickedonna Dataset (before June 2013 and after June 2016).[24]

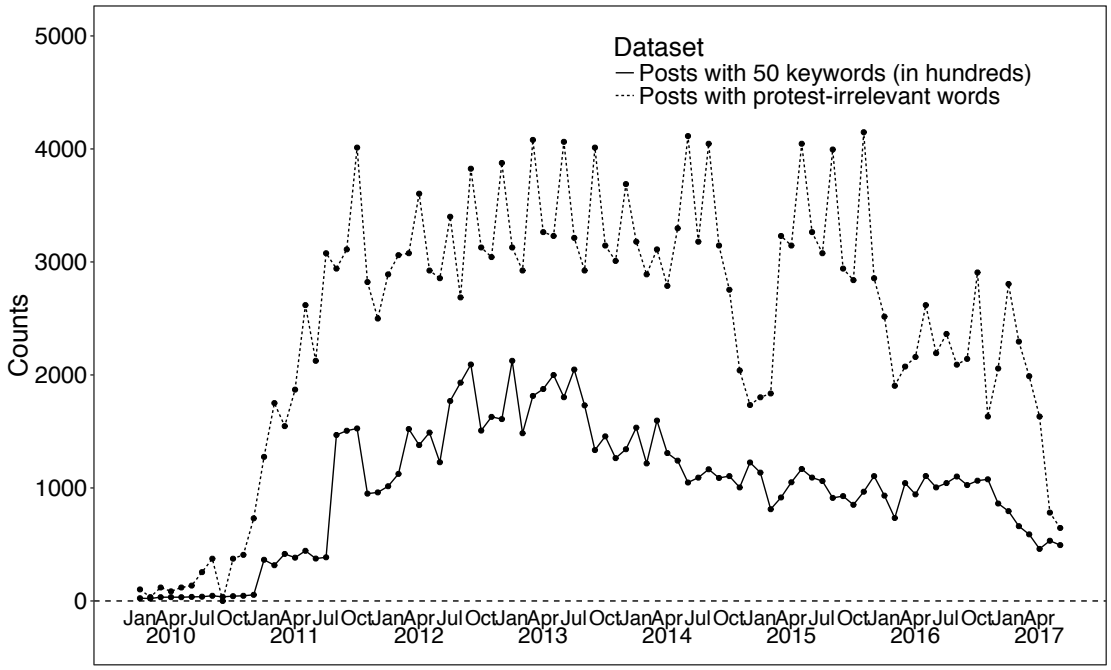
Figure 6: Number of events per month, CASM and Wickedonna. The Wickedonna Dataset is restricted to events whose sources are Weibo for a fair comparison.



Why does the Wickedonna dataset shows an increasing number of protests from 2013 to 2016 on Weibo, while CASM-China suggests a decreasing trend in the number of protests starting in 2013, as seen from Figure 6? While we do not know for sure, data suggests that the temporal trend uncovered in CASM-China better reflects the number of real-world collective action events reported on Weibo. The major reason is that Weibo reaches its height of popularity in 2012, and usage declined after 2013. We can examine the overall usage of the Weibo platform with the two placebo dataset shown in Figure 7: the number of posts over time which contain protest-related keywords T_K, and the number

of posts containing Chinese idioms (such as 三心二意, which means that a person is unable to concentrate on completing one goal at a time) that are very unlikely to be related to collective action. Figure 7 shows that placebo datasets also have declining number of

Figure 7: Number of posts per month, two placebo datasets.



posts after 2013, similar to CASM-China. The fact that the Wickedonna dataset shows the opposite trend may reflect improvements by those running Wickedonna to collect and document events.

5 Conclusion

This paper develops CASM, a new approach to identifying collective action events using social media data. In doing so, we demonstrate the advantages and limitation of using social media as a new target data source for protest event analysis, and we make methodological innovations in the creation of protest event datasets by using image as data, two-stage classification, and deep learning. We show how internal and external validation can be conducted to help make the application of computer science methods to social science

domains more practical and usable, and we discuss the ethical consideration of creating this type of system.

In this paper, we apply CASM to China, and to Weibo data. This generates one of the largest datasets of collective action in China, or any authoritarian regime. This data has a high level of spatiotemporal resolution, and spans a seven year period. We plan to make this dataset publicly available so it can be used by researchers to investigate patterns and characteristics of collective action China. Going forward, CASM can be applied to other target sources of data, covering other regions of the world.

Social media data provides unique benefits as a source of data for detecting collective action events in authoritarian regimes because it provide information when other sources such as tradition media is silent. Our intention is not to argue that social media is a better target source than traditional media or that it should replace these other target sources. Protest event analysis based on social media data should complement existing datasets in democracies, and provides a new data source for understanding patterns of collective action in authoritarian societies such as China.

In our approach, we use a two-stage classifier because social media contains grievances that fail to manifest as real-world collective action as well as descriptions of on-the-ground collective action. While this is a methodological challenge, it reflects an additional benefits of working with social media data, which has implications for theories of mobilization. Social media contains information on mobilization attempts and grievances that do not spill into the real world. Traditional media, for the most part, does not cover failures of mobilization. Research based on traditional media thus often results in selection on the dependent variable by only including successful mobilization. Social media data can better answer questions about why mobilization succeeds or fails.

Appendix A Collective Action Definition and Coding Rules

We define collective action by adapting the classic definition of McAdam et al. (2003). A collective action event is an episodic, collective event among makers of claims and their targets when:

(a) targets are political and economic power-holders (such the government);
(b) claims, if realized, affect the interests of at least one of the claimants;
(c) action of claimants is a contentious event with public physical presence

This definition, including points (a) and (b), draw directly from (McAdam et al., 2003, p. 5). By requiring the event to be episodic, we exclude regular meetings. By defining the targets of protest to include both political and economic actors, we include collective action events where the government is either a target or a mediator. For example, if a group protests against a company, but protesters turn to the government to adjudicate the dispute, the government is a mediator in the protest, not a target of the protest. We add the additional requirement of contention and public physical presence. We require the type of action to be a contentious event—boycotts, demonstrations, matches, sit-ins, strikes—instead of non-contentious events such as a fundraiser. By requiring an event to have public physical presence, we exclude events that are not visible to others, such as private group discussions or events that take place only online.

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 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.

Appendix B Selecting Keywords

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. In this section, we evaluate how our choice of K impacts data collection and the output of CASM.

The choice of K influences the data collection process because T_K expands with the size of K. While using a larger dictionary expands the coverage of protest posts, it comes at the cost of time and low specificity. Figure 8 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.

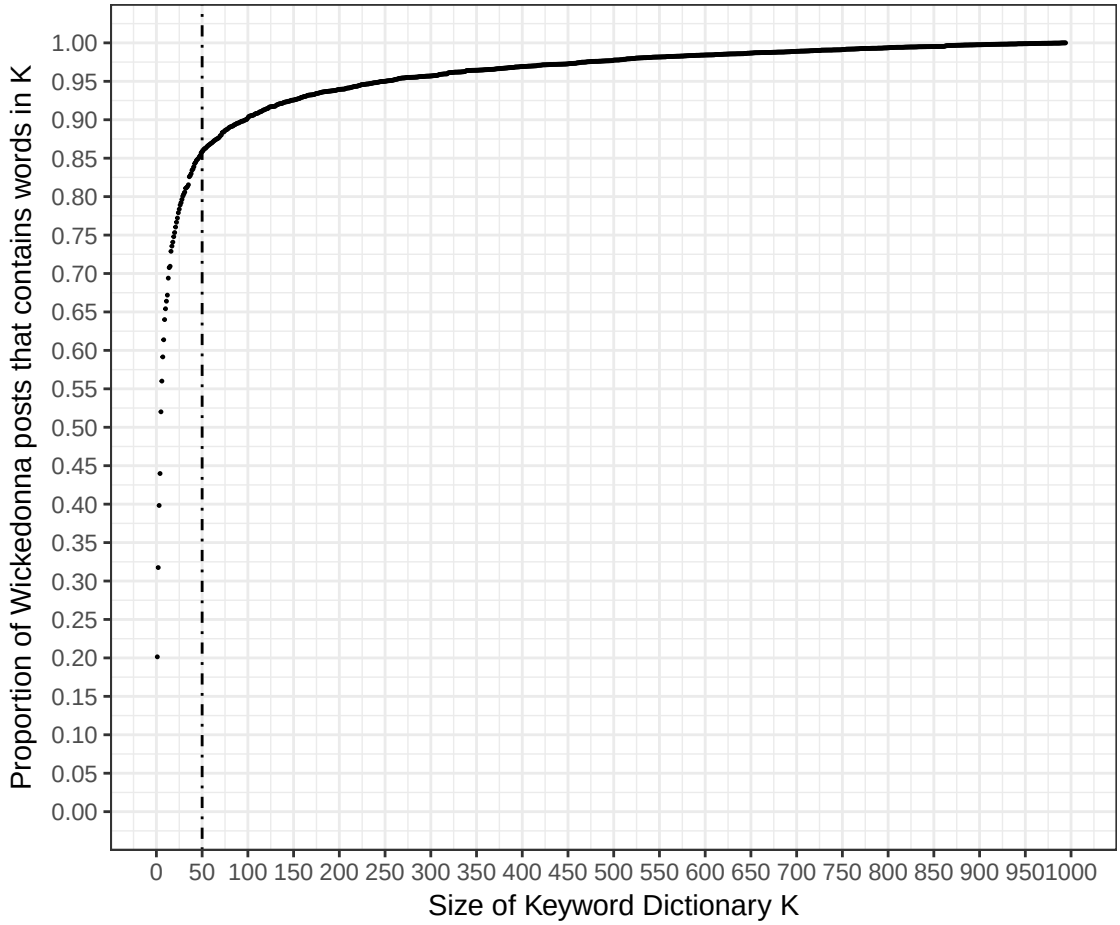


Figure 8: Coverage of protests in the Wickedonna Dataset by size of keyword dictionary; beyond 50 keywords, marginal coverage declines.

However, doubling the dictionary size will almost double the time it takes to collect posts that contain these words. Furthermore, since 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 a set of posts T_K that has lower specificity, which make 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 go a step further and 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 B plot 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

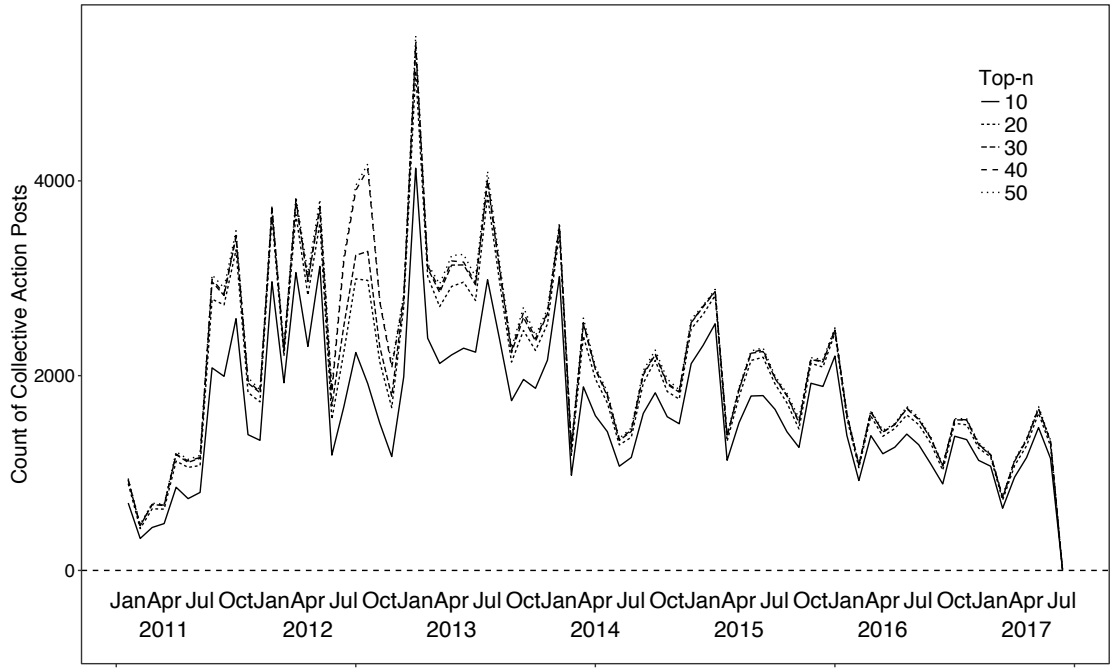


Figure 9: The number of collective-action events identified in CASM-China by the size of the dictionary.

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 of events identified. The results suggests 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.

Lastly, we show that expanding the size of dictionary leads to a decrease in our classifiers' performances. Figure 10, 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.

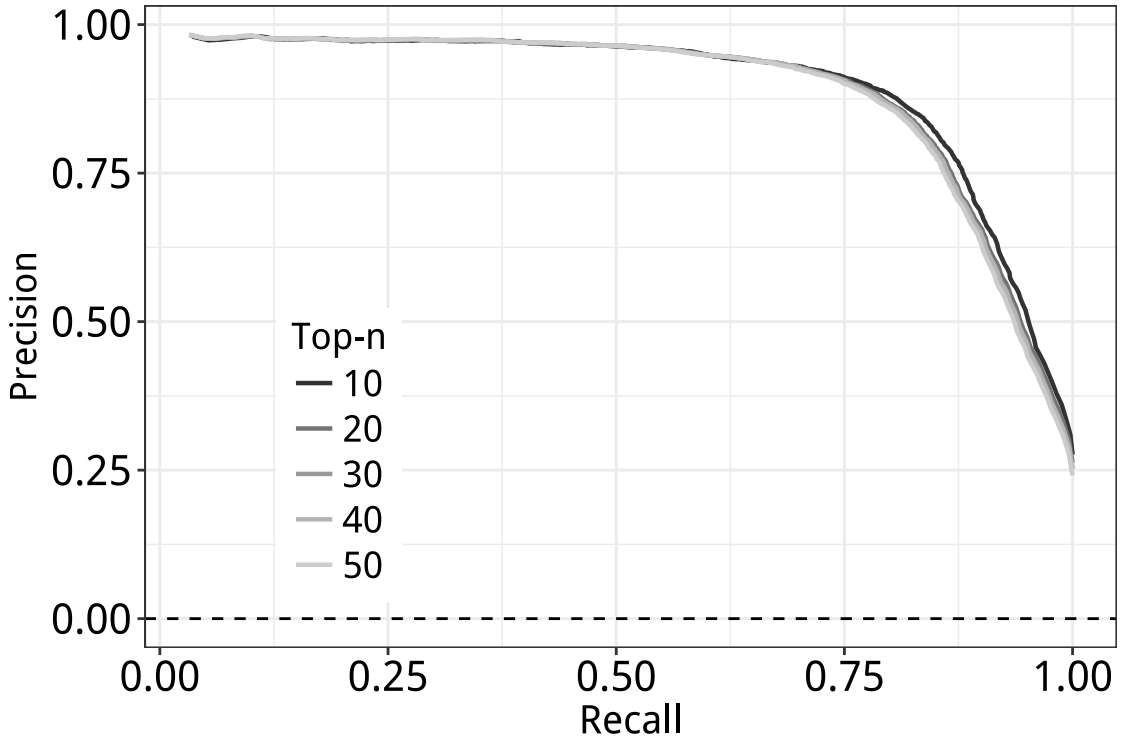


Figure 10: Precision-recall curve by size of keyword dictionary

Appendix C Generating Datasets for Comparison

In this section we describe our procedures for collecting and createing collective action events datasets in China that are used to assess the external validity of CASM-China. These datasets are shown in Table 1. For each dataset, we describes how we constructed or cleaned the data in order to compare its 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. GEDLT 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 asosciated 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 represent 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 around $\frac{27}{200} * 2214 \approx 299$ 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.[25]. 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 call 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 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.[26]

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.[27] 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.[28] 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 $\frac{17}{16276} * 264938 \approx 276$ events between January and June 2016.

Wickedonna: We introduced the Wickedonna Dataset in Section 4. 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 500 events in Wickedonna, and let human coders check whether they are in the Wickedonna dataset. We find that for 42% of the 500 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, 70% (48% + 52%* 0.42) of the events in the Wickedonna are covered by CASM-China.

For 32% of the 500 events (16.6% = 52%* 0.32 of the total population), we find that they contain words that are not in our keyword dictionary so that CASM do not collect them. 14% of the 500 events (7.3% =52%* 0.4 of the total population) are no longer available on Weibo, either due to self-deletion or censorship. The remaining posts (6.2% = 52% * 0.12) 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," 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 an Hong Kong-based NGO that aims to help labor workers bargain with employers and advocate for their rights. One

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[29] 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.

Notes

[1]Studies have identified linkages between collective action and democratization, or at least democracy-enhancing outcomes, all over the world, from Western and Eastern Europe to Sub-Saharan Africa to Latin America (Acemoglu and Robinson, 2006; Bermeo, 1997; Beissinger, 2007; Bratton and Van de Walle, 1997; Bunce, 2003; Collier, 1999; Foweraker and Landman, 2000; Geddes, 1999; Gill, 2000; McFaul, 2002; Rueschemeyer et al., 1992; Schock, 2005; Tucker, 2007; Wood, 2000, 2001). Some scholars argue that democratization is not possible without collective action (McAdam et al., 2003; Ulfelder, 2005), and other contend that individualized contention can also act as the spark to large-scale mobilization (Fu, 2017).

[2]For more on this literature, see Koopmans and Rucht (2002) and Hutter (2014).

[3]Although coined as "protest" event analysis, protest event analysis usually uses datasets that focus on specific types of collective action events of interest to scholars, e.g., racial violence, agrarian protests and rebellions, sit-ins, etc.

[4]http://ronfran.faculty.ku.edu/data/index.html

[5]Steinert-Threlkeld (2017) uses Twitter data to study mobilization networks during the Arab spring. This study is unique in that it uses social media data to study protests under authoritarian rule; however, social media data is used to shed light on the characteristics of protests, which had already been covered by traditional media, not to identify previously unknown protests.

[6]The Global Database of Events, Language, and Tone (GDELT) is a prominent example of a fully automated rule-based approach that takes pre-defined actor-verb-object phrases to find matching articles and assign them into pre-determined event categories, including protests. We empirically show the severity of coding errors of the GDELT system for identifying collective action events in Appendix C.

[7]https://bit.ly/2HrPhIj

[8]As of September 2016, Weibo had 132 million daily-active users, and 297 million monthly-active users (see http://data.weibo.com).

[9]One difference between Weibo and Twitter is that Weibo allows users to comment on a post without retweeting (similar to comments on Facebook).

[10]https://newsworthknowingcn.blogspot.com

[11]The Wickedonna dataset has strong spatiotemporal resolution and covers a wide range of issues; however, it has several limitations. First, Lu and Li never define what constitutes a protest. Some events in the dataset feature large-scale protests, while others appear to be protest by a single individual. Second, we have no information on how Lu and Li collected the events, and thus cannot ascertain what biases exist in their data. For example, it is unclear to what extent a set of keywords were used, or whether protesters would contact Lu and Li to report their protests. Both Lu and Li have been detained by the Chinese government since June 2016, and we have no way of verifying the exact procedures used to compile this data.

[12]We begin in 2010 because Weibo launched in September, 2009. The number of posts in 2010 is still very sparse, as can be seen later in Figure 6.

[13]Low likelihood of being about collective action is based on the probability assigned by a SVM classifier trained on the positive training data and first negative training data.

[14]This limited social media attention could be due to censorship or due to lack of interest. It could be due to censorship because censorship focuses on bursts of discussion but is incomplete. We examine the impact of censorship in depth in Section 4.4.

[15]We use text classifier because deep learning, SVM and Naive Bayes can take the same preprocessed text data as input. We do not compare the performance of SVM and Naive Bayes on image classifiers because they require very different preprocessing techniques on images, which is much more complex than deep learning algorithms. Scholarship also reaches a consensus that deep learning algorithms outperforms conventional algorithms by a margin (Krizhevsky et al., 2012).

[16]The precision-recall curve for the image line (dotted line) has a sharp transition point when recall = 0.5 and precision = 0.23, because only half of the posts in the validation dataset has images.

[17]The comparison between the second-stage classifier, SVM and Naive Bayes shows a similar trend.

[18]This Weibo data was shared with us by Thresher.

[19]For all comparison, we only include datasets that are open access and have event-level information instead of simple counts of events.

[20]The 12% of collective action events we do not cover are sourced by Wisenews from WeChat, another major social networking site in China. WiseNews has been indexing publicly shared articles from WeChat,

but we do not know the details of their method for obtaining WeChat data.

[21]We describe how we collected these data in Appendix C

[22]There are other human-curated protest event datasets, mostly based on newspapers, such as Cai (2010); Shao (2017). However, none of these datasets are open source and can be used to compare against our dataset.

[23]Weibo bans automated searches for words particularly like "protests" or "strikes," and for words that are very popular, such as "government", Weibo returns at most 1000 posts per search.

[24]To allow a fair comparison, the Wickedonna Dataset is restricted to events whose sources is Weibo.

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

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

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

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

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

References

Acemoglu, D. and J. Robinson (2006). Economic backwardness in political perspective. *American Political Science Review 100*(1), 115–131.

Adams, N. (2014, July). Researchers to Crowds to Algorithms: Building Large, Complex, and Transparent Databases from Text in the Age of Data Science. SSRN Scholarly Paper ID 2459325, Social Science Research Network, Rochester, NY.

Almeida, P. and M. Lichbach (2003, October). To The Internet, From The Internet: Comparative Media Coverage Of Transnational Protests. *Mobilization: An International Quarterly 8*(3), 249–272.

Azar, E. E., S. H. Cohen, T. O. Jukam, and J. M. McCormick (1972). The problem of source coverage in the use of international events data. *International Studies Quarterly 16*(3), 373–388.

Barberá, P. (2015, January). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis 23*(1), 76–91.

Beissinger, M. R. (2007). Structure and example in modular political phenomena: The diffusion of bulldozer/rose/orange/tulip revolutions. *Perspectives on Politics 5*(2), 259–276.

Bermeo, N. (1997). Myths of moderation: confrontation and conflict during democratic transitions. *Comparative Politics 29*(3), 305–322.

Bourgault, A. (2015). Freedom of the press under authoritarian regimes. *Susquehanna University Political Review 6*(1), 3.

Bratton, M. and N. Van de Walle (1997). *Democratic experiments in Africa: Regime transitions in comparative perspective*. Cambridge University Press.

Budak, C. and D. J. Watts (2015). Dissecting the spirit of gezi: Influence vs. selection in the occupy gezi movement. *Sociological Science*, 370–397.

Bunce, V. (2003). Rethinking recent democratization: Lessons from the postcommunist experience. *World politics 55*(2), 167–192.

Cai, Y. (2010). *Collective Resistance in China: Why Popular Protests Succeed or Fail*. Stanford University Press.

Chen, X. (2011). *Social protest and contentious authoritarianism in China*. Cambridge University Press.

Collier, R. B. (1999). *Paths toward democracy: The working class and elites in Western Europe and South America*. Cambridge University Press.

Crandall, J., M. Crete-Nishihata, J. Knockel, S. McKune, A. Senft, D. Tseng, and G. Wiseman (2013). Chat program censorship and surveillance in china: Tracking tom-skype and sina uc. *First Monday 18*(7).

Croicu, M. and N. B. Weidmann (2015, October). Improving the selection of news reports for event coding using ensemble classification. *Research & Politics 2*(4), 2053168015615596.

Deng, Y. and K. J. O'Brien (2013). Relational repression in china: using social ties to demobilize protesters. *The China Quarterly 215*, 533–552.

Denyer, S. (2018). Beijing bets on facial recognition in a big drive for total surveillance. *The Washington Post January 7*.

Diamond, L. (2010). Liberation technology. *Journal of Democracy 21*(3), 69–83.

Dimitrov, M. and Z. Zhang (2017). Patterns of Protest Activity in China.

Earl, J. and K. Kimport (2008, June). The Targets of Online Protest. *Information, Communication & Society 11*(4), 449–472.

Earl, J. and K. Kimport (2011). *Digitally Enabled Social Change: Activism in the Internet Age*. MIT Press.

Earl, J., A. Martin, J. D. McCarthy, and S. A. Soule (2004). The Use of Newspaper Data in the Study of Collective Action. *Annual Review of Sociology 30*, 65–80.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *The Review of Economic Studies 80*, 1422–1458.

Egorov, G. and K. Sonin (2011). Dictators and their viziers: Endogenizing the loyalty–competence trade-off. *Journal of the European Economic Association 9*(5), 903–930.

Ferdinand, P. (2000). The internet, democracy and democratization. *Democratization 7*(1), 1–17.

Foweraker, J. and T. Landman (2000). *Citizenship rights and social movements: a comparative and statistical analysis*. Oxford University Press.

Freedom House (2017). Press freedom's dark horizon. *Freedom of the Press 2017*.

Fu, D. (2017). Disguised collective action in china. *Comparative Political Studies 50*(4), 499–527.

Gamson, W. A. and A. Modigliani (1989, July). Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. *American Journal of Sociology 95*(1), 1–37.

Geddes, B. (1999). What do we know about democratization after twenty years? *Annual Review of Political Science 2*, 115–144.

Gill, G. (2000). *The Dynamics of Democratization: Elites, Civil Society and the Transition Process*. Palgrave Macmillan.

Goebel, C. (2017). Social Unrest in China A bird's eye perspective.

González-Bailón, S., J. Borge-Holthoefer, A. Rivero, and Y. Moreno (2011, December). The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports 1*.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Hanna, A. (2017). MPEDS: Automating the Generation of Protest Event Data.

Hem, M. (2014). Evading the censors: Critical journalism in authoritarian states. *Reuters Institute Fellowship Paper, University of Oxford, Trinity Term*.

Huang, R. and X. Sun (2014). Weibo network, information diffusion and implications for collective action in china. *Information, Communication & Society 17*(1), 86–104.

Hutter, S. (2014). Protest event analysis and its offspring. In D. Della Porta (Ed.), *Methodological Practices in Social Movement Research*. Oxford University Press.

Jacobs, H. (1967). To count a crowd. *Columbia Journalism Review; New York, N. Y. 6*(1), 37–40.

Jenkins, J. C. and C. M. Eckert (1986). Channeling Black Insurgency: Elite Patronage and Professional Social Movement Organizations in the Development of the Black Movement. *American Sociological Review 51*(6), 812–829.

Jenkins, J. C. and C. Perrow (1977). Insurgency of the Powerless: Farm Worker Movements (1946-1972). *American Sociological Review 42*(2), 249–268.

Kharroub, T. and O. Bas (2016, October). Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media & Society 18*(9), 1973–1992.

Kim, Y. (2014, August). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1746–1751.

King, G., P. Lam, and M. E. Roberts (2017). Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science 61*(4), 971–988.

King, G., J. Pan, and M. E. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review 107*(02), 326–343.

King, G., J. Pan, and M. E. Roberts (2014). Reverse-engineering Censorship in China: Randomized Experimentation and Participant Observation. *Science 345*(6199), 1–10.

King, G., J. Pan, and M. E. Roberts (2017). How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *American Political Science Review 111*(3), 484–501.

Koopmans, R. and D. Rucht (2002). Protest event analysis. In B. Klandermans and S. Staggenborg (Eds.), *Methods of Social Movement Research*, Volume 16, pp. 231–259. University of Minnesota Press.

Kriesi, H. (1995). *New Social Movements in Western Europe: A Comparative Analysis*. U of Minnesota Press. Google-Books-ID: Ncec7ha3pZEC.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

Lei, Y.-W. (2016). Freeing the Press: How Field Environment Explains Critical News Reporting in China. *American Journal of Sociology 122*(1), 1–48.

Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science 58*(2), 402–414.

McAdam, D. (1982). *Political process and the development of black insurgency, 1930-1970*. University of Chicago Press.

McAdam, D. and Y. Su (2002). The War at Home: Antiwar Protests and Congressional Voting, 1965 to 1973. *American Sociological Review 67*(5), 696–721.

McAdam, D., S. Tarrow, and C. Tilly (2003). *Dynamics of Contention*. Cambridge University Press.

McCarthy, J., C. McPhail, and J. Smith (1996). Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991. *American Sociological Review 61*(3), 478–499.

McFaul, M. (2002). The fourth wave of democracy and dictatorship: noncooperative transitions in the postcommunist world. *World politics 54*(2), 212–244.

McMillan, J. and P. Zoido (2004). How to subvert democracy: Montesinos in peru. *Journal of Economic Perspectives 18*(4), 69–92.

McPhail, C. and J. McCarthy (2004, August). Who Counts and How: Estimating the Size of Protests. *Contexts 3*(3), 12–18.

Miller, S. and M. J. Selgelid (2007). Ethical and philosophical consideration of the dual-use dilemma in the biological sciences. *Science and engineering ethics 13*(4), 523–580.

Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist (2011). Understanding the Demographics of Twitter Users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Volume 11, pp. 554–557. AAAI Press.

Nam, T. (2006, April). What You Use Matters: Coding Protest Data. *PS: Political Science & Politics 39*(2), 281–287.

Nardulli, P. F., S. L. Althaus, and M. Hayes (2015). A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data. *Sociological Methodology*, 1–36.

O'Brien, K. and L. Li (2006). *Rightful Resistance in Rural China*. Cambridge University Press.

O'Brien, K. and R. Stern (2007). Studying contention in contemporary china. In K. O'Brien (Ed.), *Popular Protest in China*, pp. 11–25. Harvard University Press.

Oliver, P. E. and G. M. Maney (2000). Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions. *American Journal of Sociology 106*(2), 463–505.

Oliver, P. E. and D. J. Myers (1999). How Events Enter the Public Sphere: Conflict, Location, and Sponsorship in Local Newspaper Coverage of Public Events. *American Journal of Sociology 105*(1), 38–87.

Ortiz, D., D. Myers, E. Walls, and M.-E. Diaz (2005, October). Where Do We Stand with Newspaper Data? *Mobilization: An International Quarterly 10*(3), 397–419.

Pan, J. (2015). *Buying Inertia: Preempting Social Disorder with Selective Welfare Provision in Urban China*. Ph. D. thesis, Harvard University.

Perry, E. (2002). *Challenging the Mandate of Heaven: Social Protest and State Power in China*. Armork, NY: M. E. Sharpe.

Perry, E. (2008). Permanent revolution? continuities and discontinuities in chinese protest. In K. O'Brien (Ed.), *Popular Protest in China*, pp. 205–216. Cambridge, MA: Harvard University Press.

Qin, B., D. Stromberg, and Y. Wu (2017). Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives 31*(1), 117–140.

Qin, B., D. Strömberg, and Y. Wu (2018). Media Bias in China. *American Economic Review 108*(9), 2442–2476.

Rainie, L., A. Smith, K. L. Schlozman, H. Brady, and S. Verba (2012). Social media and political engagement. *19*.

Ramakrishnan, N., P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, C. Kuhlman, A. Marathe, L. Zhao, T. Hua, F. Chen,

C. T. Lu, B. Huang, A. Srinivasan, K. Trinh, L. Getoor, G. Katz, A. Doyle, C. Ackermann, I. Zavorin, J. Ford, K. Summers, Y. Fayed, J. Arredondo, D. Gupta, and D. Mares (2014). 'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, New York, NY, USA, pp. 1799–1808. ACM.

Rasler, K. (1996). Concessions, repression, and political protest in the iranian revolution. *American Sociological Review*, 132–152.

Roberts, M. (2018). *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press.

Rucht, D., R. Koopmans, and F. Neidhardt (1999). *Acts of dissent: new developments in the study of protest*. Rowman & Littlefield.

Rueschemeyer, D., E. Stephens, and J. Stephens (1992). *Capitalist development and democracy*. University of Chicago Press.

Saraf, P. and N. Ramakrishnan (2016). EMBERS AutoGSR: Automated Coding of Civil Unrest Events. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 599–608. ACM.

Schock, K. (2005). *Unarmed insurrections: People power movements in nondemocracies*. University of Minnesota Press.

Selgelid, M. J. (2013). Dual-Use Research. In *International Encyclopedia of Ethics*. American Cancer Society.

Shao, D. (2017). The construction and application of mass incidents database in china. *China Public Administration*, 126–130.

Simonyan, K. and A. Zisserman (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the Third International Conference on Learning Representations*.

Smith, A. (2013). Civic Engagement in the Digital Age. Pew Research Center.

Steinert-Threlkeld, Z. C. (2017). Spontaneous collective action: peripheral mobilization during the arab spring. *American Political Science Review 111*(2), 379–403.

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

Sullivan, J. (2014). China's Weibo: Is faster different? *New Media & Society 16*(1), 24–37.

Tarrow, S. (2005). *The New Transnational Activism*. Cambridge University Press.

Trentham, B., S. Sokoloff, A. Tsang, and S. Neysmith (2015, July). Social media and senior citizen advocacy: an inclusive tool to resist ageism? *Politics, Groups, and Identities 3*(3), 558–571.

Tucker, J. A. (2007). Enough! electoral fraud, collective action problems, and post-communist colored revolutions. *Perspectives on Politics 5*(3), 535–551.

Tufekci, Z. and C. Wilson (2012). Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of Communication 62*(2), 363–379.

Ulfelder, J. (2005). Contentious collective action and the breakdown of authoritarian regimes. *International Political Science Review 26*(3), 311–334.

Ward, M., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford (2013, January). Comparing GDELT and ICEWS event data. *Analysis 21*, 267–297.

Wong, E. (2012). China's growth slows, and its political model shows limits. *The New York Times May*(10).

Wood, E. (2001). An insurgent path to democracy: popular mobilization, economic interests, and regime transition in south africa and el salvador. *Comparative Political Studies 34*(8), 862–888.

Wood, E. J. (2000). *Forging democracy from below: Insurgent transitions in South Africa and El Salvador*. Cambridge University Press.

Yang, G. (2003). The internet and the rise of a transnational chinese cultural sphere. *Media, Culture & Society 25*(4), 469–490.