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

Han Zhang* Jennifer Pan†

January 15, 2019

Abstract

Protest event analysis is an important method for the study of collective action and social movements, which typically draws on traditional media reports as the data source. We introduce Collective Action from Social Media (CASM)—a system that uses convolutional neural networks on image data and recurrent neural networks with long short-term memory on text data in a two-stage classifier to identify collective action events occurring offline. We implement CASM on Chinese social media data and identify 142,427 collective action events from 2010 to 2017 (CASM-China). We extensively evaluate the performance of CASM through cross-validation, out-of-sample validation, and comparisons with other Chinese protest datasets. We assess the impact of online censorship, and find that it does not substantially limit our identification of events. Compared to other datasets of protests, CASM-China identifies relatively more rural, land-related protests, but identifies few collective action events related to ethnic and religious conflict.

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

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

1 Introduction

Protest event analysis is an important method for social movement research (Hutter, 2014a; Koopmans and Rucht, 2002; Olzak, 1989; Earl et al., 2004) which has played a key role in the development of political process theory (Jenkins and Perrow, 1977; McAdam, 1982), theories of resource mobilization (Jenkins and Eckert, 1986), study of new social movements in Europe (Kriesi, 1995), and comparative studies of global and transnational activism (Tarrow, 2005). Protest event analysis requires the creation of event datasets that allow researchers to systematically assess the occurrences and features of collective action events across geographic boundaries and over time.

The main data source, or target source, for the creation of collective action datasets for protest event analysis 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, and allow researchers to quantitatively assess both the occurrence of these events across geographic boundaries and over time, as well as the features and characteristics of these events. Well-known examples of newspaper-based collective action datasets include the US-focused Dynamics of Collective Action (DoCA), which draws from *The New York Times* reporting between 1960 and 1995 (McAdam and Su, 2002); the PRODAT project, which used German newspapers from 1950-2001 (Rucht et al., 1999); and the European Protest and Coercion Data, which identified events based on newspapers in 30 European countries.[1]

Despite its many benefits, biases in media coverage of collective action events limit the use of traditional media for protest event analysis (McCarthy et al., 1996a; Ortiz et al., 2005). Newspapers are more likely to report on larger protests, and on protests that are more sensational in nature. Certain news outlets are more likely to report on some types of protest than other types. Research shows that selection bias in newspaper coverage of

protests can lead to bias in datasets that are 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, Google search records, and government archives (Almeida and Lichbach, 2003; Earl and Kimport, 2008; Gamson and Modigliani, 1989; McCarthy et al., 1996b).

It is particularly challenging to use traditional media as a target source to study collective action in authoritarian regimes. This is unfortunate because collective action is especially important for our understanding of social, political, and economic processes in these countries where opportunities for many forms of expression and representation are limited. Independent measures of collective action would be highly valuable for numerous scientific and public policy purposes, but authoritarian regimes impose strict controls on news reporting through state ownership of media outlets (Egorov and Sonin, 2011; Qin et al., 2012; Stockmann, 2013), through repression and co-optation of private media outlet owners (McMillan and Zoido, 2004), and through intimidation and surveillance of domestic and foreign journalists (Bourgault, 2015; Freedom House, 2017; Hem, 2014). As a consequence, many collective action and protest events that happen in authoritarian regimes are not reported in traditional media, either by local or foreign news outlets, and answering even basic factual questions about collective action events is a challenge.

The adoption of digital technologies provides new opportunities for scholars to learn about collective action and to complement what we already know about collective action from traditional media reporting. The internet, social media, and 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 it 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. Researchers have already used social media to study substantive topics in contentious politics and social movements (Budak and Watts, 2015; Barberá, 2015; González-Bailón et al., 2011; Steinert-Threlkeld et al., 2015; Steinert-Threlkeld, 2017), but the digital traces left by protesters, bystanders, and commentators also provide us with new ways of identifying collective action events.

In this paper, we create CASM (Collective Action from Social Media)—a system that uses social media data to identify collective action events occurring in the real world. CASM is intended to identify events that happen outside of the internet, that have public physical presence; CASM is not focused on identifying online mobilization or online collective action (Bennett and Segerberg, 2012; Castells, 2015; Romero et al., 2011). In this paper, whenever we refer to collective action events, we are referring to offline events.

CASM identifies collective action events from social media posts by applying deep learning algorithms, using image and textual data, in a two-stage classifier to identify posts about collective action events. CASM uses convolutional neural network (CNN) for image classification, and a combination of convolutional and recurrent neural networks with long short-term memory (CNN-RNN) for textual analysis. These deep learning algorithms jointly model the data representation (how to represent raw data as features) and perform classification, and these algorithms allow for transfer learning, reusing models based on large datasets as a starting point for our task of identifying collective action events. We use these deep learning algorithms in a two-stage classifier, which allows us to overcome the challenge of distinguishing between social media posts that describe offline collective action events from posts that discuss similar topics but do not manifest as offline collective action events. We test our system through extensive internal and external validation, which we hope offers a template for how computer science methods can be made more practical

and usable for social science research.

We implement CASM for China (CASM-China) using social media data from Sina Weibo, a popular Chinese microblogging platform, and identify 142,427 events from January 1, 2010 to June 30, 2017, with events found in over 96% of counties in China. 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-China. In assessing the external validity of CASM-China, we find that the system will miss collective action events taking places in ethnic 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.

We proceed in five sections. Section 2 describes the advantages and limitations of using social media as a target data source for identifying collective action events. In Section 3, we describe the details of CASM and its implementation on Chinese social media data. We define collective action, discuss how we collect and preprocess data, describe the architecture of the CNN and CNN-RNN models, detail how these models are trained, describe performance of the first-stage and second-stage classifiers, and show how we identify unique events from posts. Section 4 presents a description of the output of CASM for China, a dataset we call CASM-China, along with an assessment of its external validity through comparison with other event datasets and evaluation of the impact of censorship. Section 5 discusses how CASM can be implemented beyond China, and Section 6 concludes.

2 Social Media as a Target Data Source

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. The characteristics of social media data that provide unique advantages for protest event analysis include: 1) scale, 2) unmediated channel, and 3) diversity.

Scale: More than half of the world's population are on the internet, and social media is used in every country with internet access (Rainie et al., 2012).[2] The scale of social media vastly exceeds that of traditional media—an average of 31 million messages are sent every minute on Facebook; an average of nearly 350,000 tweets are made every minute.[3] Social media gives every individual the power to broadcast, and even if only a small minority of social media users talk about offline collective action, the number of collective action events reported on social media will still vastly outstrip what can be reported by traditional media sources.

Unmediated channel: From the Arab Spring to Occupy to the MeToo movement, social media has become an important venue for protesters to speak out (González-Bailón et al., 2011; Budak and Watts, 2015; Barberá, 2015; Steinert-Threlkeld et al., 2015; Steinert-Threlkeld, 2017). Thus, 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 direct understanding of the grievances, problems, and issues that mobilize, rather than one mediated by news organizations (Koopmans, 2004).

Diversity: Social media data also gives researchers access to a more diverse set of collective action events, including those widely varying in scale. Traditional media are more likely to report on larger protests, and on protests that are more sensational in nature. Individuals on social media will no doubt talk about large-scale protests, but they will also report small-scale and medium-scale protests, and they may report collective action events on social media that are not violent or shocking.

These characteristics of using social media allow us to detect events that otherwise might have gone unnoticed, and to learn about collective action from the perspective of protesters. This advantage is especially crucial in authoritarian regimes where social media has become an important channel for dissent when traditional media is silent (Smith, 2013; Trentham et al., 2015; Yang, 2003).

However, there are also characteristics of social media data that can generate biases, gaps, and errors in social media-based datasets: 1) non-representativeness, 2) online censorship, 3) fast-paced technology change, and 4) brevity of content.

Non-representativeness: 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). We know that users of social media platforms constitute a non-random sample of the population (Mislove et al., 2011), which means individuals who use social media to talk about offline collective action may not be representative of everyone who engages in offline protest. For example, we would identify few collective action events from social media data in countries such as Iraq, Libya, or Turkmenistan because of low social media penetration, and protesters are unlikely to use social media to talk about their activities. We may identify more collective action events involving younger, wealthier, more educated, more urban protesters who have higher rates of adoption of social media. How protesters who post on social media compare to the overall population of

protesters will vary by country. The bias of social media toward younger people might be less problematic for identifying collective action events from online data in a country such as Saudi Arabia where nearly 50% of the population is under 25 than in a country such as Germany or Japan where less than 25% of the population is under 25.[4]

Online censorship: Social media is subject to censorship, especially in authoritarian regimes, which use a range of strategies to limit online expression. These range from blocking users in the country from accessing certain websites (e.g., China's Great Firewall, Iran's Intranet) (Deibert, 2008) to filtering search results (Bamman et al., 2012) to removing content after it has appeared online (King et al., 2013; Zhu et al., 2013) to using physical repression to induce self-censorship (Pan and Siegel, 2018; Stern and Hassid, 2012). As a result of government censorship strategies, individuals may self-censor and avoid discussions of collective action online; individuals who try to express themselves may be unable to do so, and individuals in general may be less likely to engage in collective action because the diffusion of information about these events is constrained. In addition, even if protesters talk about collective action on social media, governments can make it difficult for scholars to systematically gather social media data about collective action.

Fast-paced technology change: Social media changes rapidly. Topics of discussion change. Language and norms are fluid. Social media platforms routinely change their features and algorithms, which can change what data is available and over time comparisons. In addition, new social media platforms can emerge to displace existing platforms. This means that more collective events may be detected in social media data gathered soon after it was made public than in social media data gathered long after it was posted. This also means that social media data may not be available for long periods of time, depending on the lifecycle of social media platforms.

Brevity of content: Social media messages are often short. When discussing offline collective action, key pieces of information that would always appear in a news article (e.g., who, when, what, where, how) may be missing. This means detailed information on the features and characteristics of protest may not always be available.

These characteristics of social media will generate biases, gaps, and errors in protest event data based on social media. Scholars using these data should be mindful of these limitations, but they should not prevent us from using social media as a source of data for protest event analysis because this approach does allows us to identify numerous collective action events that might otherwise have escaped notice.

3 CASM: Collective Action from Social Media

In this section, we describe our system for identifying collective action events. Before delving into the system, it is important to clarify what we mean by collective action. We draw from (McAdam et al., 2003, 5) and define collective action as an episodic, collective event among makers of claims and their targets when:

(a) targets are political and economic power-holders (such as 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 involving three or more people.

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. By requiring the type of action to be contentious—boycotts, demonstrations, marches, sit-ins, strikes—we exclude 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. By requiring at least three people, we are setting a low threshold.[5] This definition of collective action is similar to classical protest event studies in that our primary focus is on identifying events. This definition is also related to the concept of contentious performance: "learned and historically grounded ways of making claims" (Tilly, 2008, 4) because we focus on a subset of contentious performances, but our definition differs from the theoretical focus of contentious performance because our primary aim is not to capture the varied ways in which claims can be made.

3.1 Collecting and Preprocessing Social Media Data

We use social media data from Sina Weibo (hereafter Weibo), China's biggest microblogging platform.[6] Weibo allows messages up to a maximum 140 characters. Users can mention or talk to other users, use hashtags, follow other users, and repost. Weibo is like Twitter in that it is an open platform where users do not have to follow another user to read their posts.[7]

The quantity of social media posts is vast, and in relation to the universe of social media posts, posts containing discussions of collective action events are extremely rare. Out of a random sample of 20,000 geo-coded posts from Weibo, we identified one post discussing a real-world collective action event. This implies that less than 0.01% of social media posts in China discuss protest. Thus, instead of collecting a random sample of all posts, we collect posts, T_K, that contain one or more keywords (K) related to collective action. Note that most posts T_K will not relate to real-world collective action events.[8] For example, the term "protest" (抗议) is the most frequent keyword in our keyword set for China, but posts containing this keyword such as "My stomach is protesting I'm so hungry," "I wish Chinese people had the same right to protest as people in democratic countries," and

"The US government should focus on their own protests first before paying attention to the protests in China" do not meet our definition of collective action.

The set of keywords K used to collect posts T_K can be curated by experts, or it can be calculated by identifying frequently occurring and/or differentiating keywords from social media posts known to discuss collective action. We create the set of protest-related keywords K 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 gathered a daily list of protests in China from social media reports on Sina Weibo, Tencent Weibo, Qzone, and other online platforms, and published this list on their blog.[9] Each protest is associated with a number of related social media texts, images, and sometimes videos. In total, the Wickedonna Dataset contains 67,502 protests described by 240,521 text-based posts and 233,288 images and videos. The Wickedonna Dataset has strong spatiotemporal resolution but we do not know Lu and Li's methodology for gathering these data or their criteria for inclusion.[10] We chose the 50 most frequently occurring words, excluding stopwords,[11] from the Wickedonna Dataset 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 (see Supplementary Appendix for our validation of the size of K).

We collected all Weibo posts published between January 1, 2010 and June 30, 2017 that contain at least one of the words in K.[12] Our set of posts T_K includes approximately 9.5 million posts from Weibo.[13] For each post, we collect the text, images (if there are any), as well as available meta data of the post such as the time of posting, the number of reposts, and the latitude/longitude of the post (when the account has geo-location enabled).

Chinese text does not require preprocessing steps of stemming or lowercasing common to English-language data. Instead, Chinese text is presented without whitespaces, so we preprocess posts by segmenting characters to delineate words. Our segmentation al-

gorithm, Jieba, uses a preset dictionary structure to support word graph scanning.[14] We use the largest dictionary available for Jieba and add in approximately 1,000 frequently words (excluding stopwords) from the Wickedonna Dataset. The segmenter builds a directed acyclic graph for all possible word combinations, and uses a Hidden Markov model with the Viterbi algorithm to identify words. Because we are using deep learning models, it is not absolutely necessary to conduct word segmentation; however, we segment because incorporating well-defined boundaries of the text (here, words) helps accelerate models' feature learning process. We remove punctuation and only keep posts that have at least eight segmented words. Among these retained posts, we remove stopwords and emojis.[15]

For images, the default upload file format on Weibo is JPEG. We keep all JPEG files, and exclude GIF files, which represent less than 1% of the images we encountered. Each JPEG file is rescaled to an 100×100 pixel image in color, which means image files are represented as arrays where three values—for red, green, and blue (RGB)—are associated with each pixel.

3.2 Identifying Collective Action Posts

Existing methods of building collective action datasets have used human coding and automated rule-based approaches, which fall short when dealing with social media data. The scale of social media data makes human coding impractical when the goal is to capture overall trends rather than study specific cases. The brevity and changing nature of social media posts—in terms of language, style, slang—challenge automated rule-based approaches, which often rely on the applicability of pre-defined rules (based on either keywords, parts of speech tagging or pre-defined grammatical phrases) to find matching content (Saraf and Ramakrishnan, 2016).[16]

We use supervised machine learning algorithms where humans code training data and

algorithms are "trained" with this human-coded data to create a collective action event dataset. Supervised-learning approaches are more adaptive to different data sources and more flexible than rule-based approaches (Nardulli et al., 2015a; Hanna, 2017; Croicu and Weidmann, 2015).[17]

Specifically, we use deep learning algorithms in two-stage classification to identify posts related to collective action, which we call $T_{protest}$. Deep learning algorithms are a class of machine learning algorithms based on the framework of artificial neutral networks (Bengio et al., 2015; LeCun et al., 2015). Deep learning algorithms have helped make significant advances in many machine learning tasks, especially tasks related to the analysis of images and text such as image classification (He et al., 2016a; Simonyan and Zisserman, 2014), multiple object detection (Ren et al., 2015), automated image captioning (Shin et al., 2016), voice recognition (Hinton et al., 2012; Dahl et al., 2012), machine translation (Bahdanau et al., 2014; Sutskever et al., 2014), and parts of speech tagging (Santos and Zadrozny, 2014). Deep learning algorithms are just beginning to be used in social science research, with some emerging work analyzing image data (Torres, 2018; Won et al., 2017). Our work expands on this emerging strand of social science research by using deep learning for image and textual classification.

Deep learning algorithms differ from conventional machine classification methods in two main ways. First, conventional machine classification methods require users to decide how to transform data from their raw form (pixel values in images, words in documents) to numerical representations (sub-regions of images relevant to a specific problem, vector of count of words). In contrast, how data should be transformed into numerical values is modeled in deep learning algorithms that "discover" optimal data representation for a classification task. Second, deep learning algorithms allow for transfer learning, where a model developed for one task can be reused as the starting point for a model on a different task. For instance, transfer learning can boost performance when training data available for

a specific task is small.

Below, we describe the convolutional neural network (CNN) we use for image classification and the combined convolutional and recurrent neural network with long short-term memory (CNN-RNN) we use for text classification.[18] We use the same CNN and CNN-RNN classifiers—in terms of architecture, transfer learning, and training method—in the first- and second-stage classifiers. We then describe the first-stage classifier, followed by the second-stage classifier.

3.2.1 Convolutional Neural Network for Image Classification

We use a CNN for image classification. A CNN is a model that consists of a series of operations, called layers, where each layer takes the output of the previous layer as input, and after performing some operation on it, passes the output to the next layer. CNNs get their name from the operation of convolution—element-wise multiplication between matrices followed by summation. The output of convolution is called a "feature map," and each layer can contain multiple feature maps. After each convolution, an activation function is applied to introduce non-linearity, producing a "rectified feature map" because most real-world data is non-linear yet convolution is a linear operation. The most common activation function for CNNs is ReLU (Rectified Linear Unit), an element-wise operation applied per pixel to replace all negative pixel values with zero. Feature maps and rectified feature maps are high dimensional, so spatial pooling is often applied to reduce dimensionality. Max pooling, taking the largest element from a rectified feature map within a defined window, is most commonly used, but average pooling and other forms of pooling can also be applied. Finally, a dropout layer, which is a regularization technique, occurs before subsequent convolution operations to reduce overfitting. Generally, multiple convolution layers are used to extract useful features from the raw data. After these layers, one or more "fully connected layers," often a type of neural network called a multilayer perceptron, learns

non-linear combinations of the features generated from the convolutional layers and uses all features to classify the image.[19] The final fully connected layer typically uses a softmax activation function, or normalized exponential function, to generate the output value.

There are many different variants of CNNs, which differ based on the architecture of the network as well as parameters such as the number of filters, the filter size, and stride. The architecture we use is called VGGNet, also known as VGG or VGG-16, which we chose based on its conceptual simplicity, ease of implementation, and wide-ranging applications (Simonyan and Zisserman, 2015).[20] VGGNet uses 16 convolutional layers to extract features, and three fully connected layers to perform classification. VGGNet uses small filter sizes (3×3) and more layers (16), instead of larger filter sizes (7×7) and fewer layers, as was common in previous models (Krizhevsky et al., 2012).[21] In VGGNet, a ReLU operation follows each convolutional layer, and max pooling is performed after the second, fourth, seventh, tenth, and thirteenth convolutional layers.

VGGNet was trained on a set of 1.2 million images, classified into 1000 categories. We do not use the entire pre-trained VGG model because human faces and crowds (not to mention collective action events) are not included among the 1000 categories VGGNet was originally trained for. Instead, we train and fine tune the last four convolutional layers with our own data, as illustrated in Figure 1.[22] We do not change the first 12 convolutional layers of VGGNet because they identify more basic features of images (e.g., edges, circles), while subsequent layers use these basic features to learn more complex features (e.g., human faces, signs, and placards) specific to our task. The structure of the last four layers we use is consistent with the original VGG architecture in terms of filter size and stride. After the convolutional layers, we added a fully connected layer with ReLu activation and dropouts, and a second fully connected layer with logistic sigmoid function to output the binary-class probability.[23] The final CNN output is a probability between 0 and 1, where 1 means the image is certain to represent offline collective action and 0 means the image does not

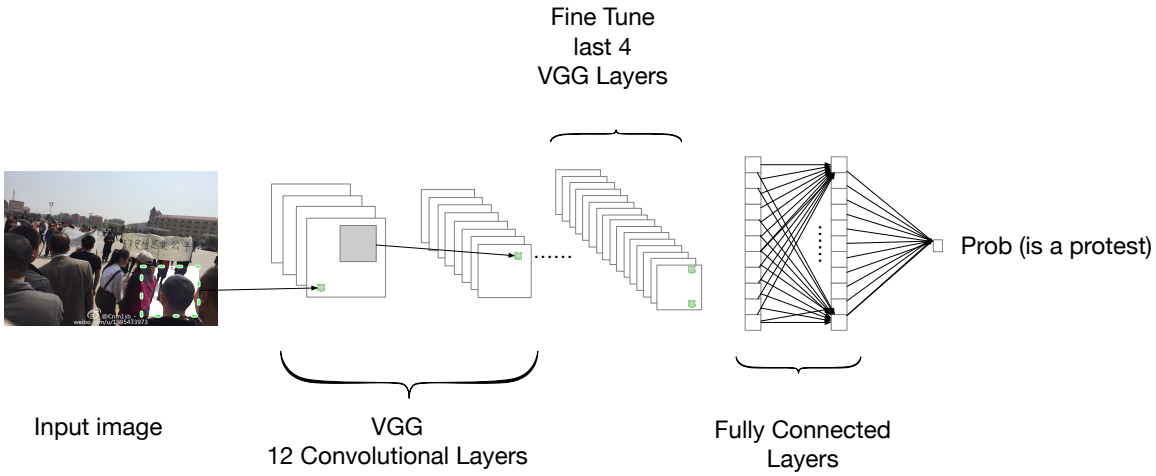


Figure 1: Illustration of our CNN architecture for image classification.

represent offline collective action.

We train this model to minimize cross entropy loss because our output is binary:

$$L(p, y) = -\frac{1}{N}\sum_{n=1}^{N}[y_n(log(p_n) + (1 - y_n)log(1 - p_n)] \quad (1)$$

where p is the output of predicted probability, y is the labels of the training data, and N is the number of images; p_n is the nth output, and y_n is the nth ground truth label. We minimize cross entropy loss by using an adaptive gradient-based optimization algorithm (Kingma and Ba, 2014).[24]

3.2.2 Convolutional and Recurrent Neural Network for Text Classification

To classify our text data, we use a model that combines convolutional layers with a recurrent neural network (RNN) with long short-term memory (LSTM) architecture. Recurrent neural networks are used extensively in dealing with sequential data, and have set the standard for performance on natural language processing tasks such as speech recognition and machine translation (Bahdanau et al., 2014; Mikolov et al., 2010; Sak et al., 2014). RNNs are a type of model that performs the same operation repeatedly on sets of sequential in-

puts. Central to a RNN is a state vector that accepts an input and the previous state to produce a new state and output. The shortcoming of "vanilla" RNNs is that they are difficult to optimize due to the effect of vanishing gradients (Pascanu et al., 2013). LSTMs were created as a special kind of RNN that can learn long-term dependencies in a computationally tractable manner with "cell" vectors that control what information from the previous sequence of operations is retained (Hochreiter and Schmidhuber, 1997).

The architecture we use combines an embedding layer, convolutional layers, a LSTM layer, and two fully connected layers (Sainath et al., 2015; Zhou et al., 2015; Xiao and Cho, 2016; Wang et al., 2016b). This architecture is shown in Figure 2.

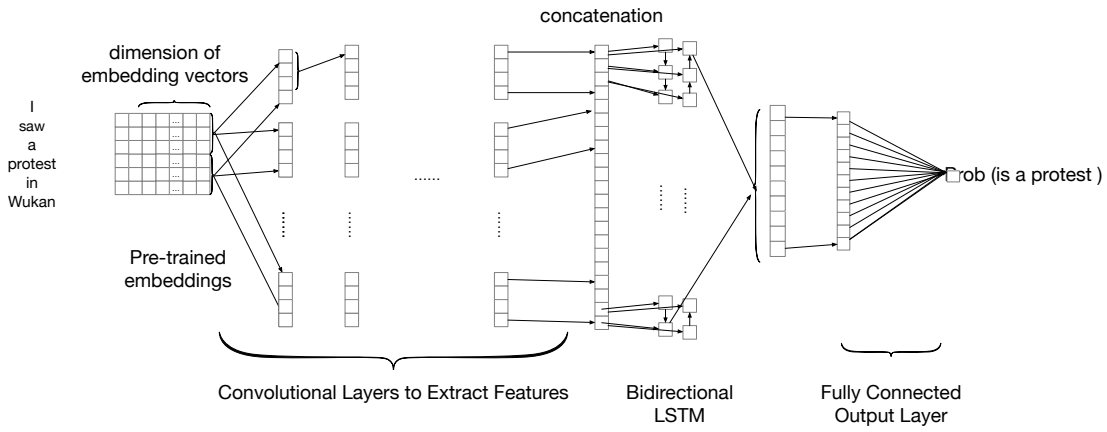


Figure 2: Illustration of our CNN-RNN architecture for text classification.

The embedding layer, shown on the left of Figure 2, is an operation to provide each word in our text data with a dense representation of the word and its relative meanings (Peters et al., 2018). Although social media text are short, the set of words, or vocabulary, used on social media is large. This means that when social media texts are represented as a "bag-of-words"—an unordered set of words—not only is word ordering lost but the vector that represents the text is exceptionally sparse. With word embeddings, words are represented by dense vectors, where a vector represents the projection of the word into a continuous vector space, and the position of the word in that space is learned based on

other words that surround the word in text. In our case, each word is represented by a 128-dimension word vector we trained with a continuous skip-gram model using 20 million Weibo posts (Mikolov et al., 2013).[25] This embedding layer can be thought of as a type of transfer learning, where we use the word vector model to add information about each word in our data.

The convolutional layers in the CNN-RNN extract features just like the convolutional layers of the CNN.[26] For image analysis, the matrices where convolution is performed are sub-regions of the image's pixels. For text analysis, the matrices are sub-regions of a $m \times n$ matrix, where m is the number of words in the social media post, and n is 128 because each word is represented by a vector of length 128. The very left of Figure 2 shows a short text: "I saw a protest in Wukan," which is represented as a matrix of size 6×128. As with image data, sub-regions are defined by filter size and stride.[27] Here, we also use a ReLu activation function, add max pooling after each convolution layer, and apply dropout.

Instead of using convolutional layers for feature extraction, we could have defined features ourselves, such as using bag-of words as features or n-grams as features. We use convolutional layers for feature extraction rather than use bag-of-words features to avoid the loss of word order and information about grammatical syntax. For example, CNNs could capture features such as the phrase "defend my rights" (维权) regardless of where or how it appears in social media posts—e.g., "I protest to defend my rights," "Why should I defend my rights?," or "Company X can defend my rights as a consumer." We use convolutional layers rather than n-grams because n-grams exponentially increase the size of the vocabulary, introducing a high level of noise (Tan et al., 2002).

We use LSTM on top of the convolutional layers because LSTMs perform better in preserving long-range dependencies within sentences and short texts (Sutskever et al., 2014). Long-range dependencies matter because meaning in a sentence or social media post is often determined by words that are not very close together. For example, a social media post

such as "The people in the square were wearing ponchos during the protest because of the heavy rain" is about people protesting not about people wearing ponchos, and LSTM can capture the long-range dependence between people and protest. Our LSTM layer is fixed to a bidirectional LSTM, which scans the inputs in forward and reserve order, preserving the proceeding and following features (Schuster and Paliwal, 1997).[28]

Finally, similar to the CNN model, there are two fully connected layers, where the first applies the ReLU activation function and dropouts, and the second is a logistic sigmoid function that outputs the binary-class probability of whether the post's text is discussing real-world collective action. To train this model, we again minimize cross entropy loss with an adaptive gradient-based optimization algorithm.

3.2.3 First-Stage Classifier

The first-stage classifier uses the CNN model described in Section 3.2.1 to classify image data and the CNN-RNN classifier described in Section 3.2.2 to classify text data. In this first-stage, we train the CNN model using a random sample of 230,000 images[29] from the Wickedonna Dataset as our positive training data (examples of images that pertain to collective action events). We use a random sample of 230,000 images from geo-located Weibo posts as the negative training data (examples of images that do not relate to collective action).[30] We train the CNN-RNN model using the 240,521 text-based posts from the Wickedonna Dataset as the positive training data. We use a random sample of approximately 200,000 geo-located posts from Weibo as the first negative training data.[31] This random sample of posts are extremely unlikely to contain discussions of collective action events, and contain few of the keywords from K. If we only use these data as the negative training data, then the positive training data would all contain protest-related words while the negative training data would not, which biases the classifier into making predictions about collective action based on whether or not a protest-related word is present. To ame-

liorate this issue, we also use approximately 450,000 posts that contain keywords from K but that have very low likelihood of being about collective action as the second negative training dataset.[32]

After training the CNN and CNN-RNN models with these data, we use the trained models to make predictions of the approximately 9.5 million Weibo posts containing at least one of the words in K. If the input Weibo post only contains text, then the CNN-RNN model generates the predicted probability that the text relates to offline collective action (p_{text}). If the input post contains text and images, the CNN-RNN model generates the predicted probability that the text related to offline collective action (p_{text}), and the CNN model assigns the predicted probability for each associated image. There are just under 3.6 million images in the 9.5 million posts. When there are multiple images associated with a social media post, we take the largest predicted probability as p_{image}.

Figure 3 shows 11 images whose predicted probabilities as assigned by the CNN in the first-stage range from 0 to 1.0. The images contained in Figure 3 are those whose predicted probabilities are closest to the integer values listed in the figure. For example, the third image from the left (a night street scene) is the image with predicted probability closest to 0.3. Figure 3 suggests that the image classifier has construct validity. Images with higher predicted probabilities of relating to collective action are images that contain crowds, signs, and placards with text, and government buildings. The appearance of a picture containing text (third image from the right, with predicted probability of 0.8) is also reassuring since Chinese social media users often post images of text to discuss sensitive topics in an attempt to avoid censorship.

When a Weibo post contains both p_{text} and p_{image}, we combine the two predicted probabilities to obtain a single predicted probability for each post. How the probabilities are combined depends on two tuning parameters: α and β. α controls how much information we should borrow from text versus the image. If α is higher, more weight is placed on

Figure 3: Images with their predicted probabilities of relating to collective action from CNN in the first-stage classifier; some images are cropped for presentational purposes.

the output of the image classifier. β controls how much extra up-weight we should give to posts that contain both text and images. The intuition here is that the existence of images in the social media post can be informative. Of 10,000 human-coded posts that contain protest-related words,[33] only 23.9% contain images in addition to text, but among posts coded as related to collective action in the test data, 56.9% contained images in addition to text, which suggests that protesters may post pictures strategically when publicizing their efforts on social media.

Formally, the predicted probability of the post is given by the following equation:

$$p = \begin{cases} \frac{p_{text} + \alpha \cdot p_{image}}{1+\alpha} \cdot \beta & \text{if the post has images} \\ p_{text}, & \text{otherwise} \end{cases} \tag{2}$$

We use cross-validation to select the optimal values of α and β.[34] The optimal α and β for the first-stage classifier is 0.33 and 1.10, respectively. The α is smaller than 1, which means that relatively more information is extracted from the CNN-RNN model of text, but β is larger than 1, confirming our intuition that a post with both text and images is more likely to be about collective action than those with only text.

We evaluate the performance of the first-stage classifier with cross-validation and out-of-sample validation. Cross-validation is the dominant approach for evaluating machine learning systems of event detection (Nardulli et al., 2015b; Hanna, 2017). The training 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. The first-stage classifier performs extremely well in cross-validation, with a maximum F_1 score of 0.96 (precision = 0.95, recall = 0.96).[35] The first-stage classifier vastly out-performs

random guess, and above the range of F_1 scores (0.6-0.8) for existing systems (Hanna, 2017; Adams, 2014).

The problem with cross-validation is that 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 in China could be based on a definition of collective action that differed from ours, and draws from a broader range of data sources. Therefore, precision and recall based on cross-validation can paint a rosier picture of the algorithm performance than is warranted. To address this problem, we conduct out-of-sample validation to mimic the context where the classifier will be used, thus providing a more realistic evaluation of the system. We take a stratified random sample of 200 posts per each keyword among from the 9.5 million posts collected between 2010 and 2017 (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 collective action event or not per our definition.[36] Then, we assess the performance of our classifier based on this independent validation set. The first-stage classifier achieves maximum F_1 score of 0.70 (precision = 0.63, recall = 0.79). Figure 4 shows the precision-recall curve of the CNN image classifier alone, the CNN-RNN text classifier alone, and the combined classifier from the first-stage based on out-of-sample validation.[37] There is often a trade-off between precision and recall, but we can see from this figure that the text-based classifier outperforms the image classifier, and the combined classifier outperforms both across the precision-recall curve.[38] The first-stage classifier correctly classifie posts about collective action as collective action—Figure 5 shows a post about collective action that is identified by the first-stage classifier as such (true positive). The first-stage also correctly classifies posts containing words K that are not related to collective action—Figure 6 shows a post unrelated to collective action that is classified at not relating to collective action (true negative). Not that the post in Figure 6 appears in our data because it contains the word "surrounded by a mob" (围堵), which is in K.

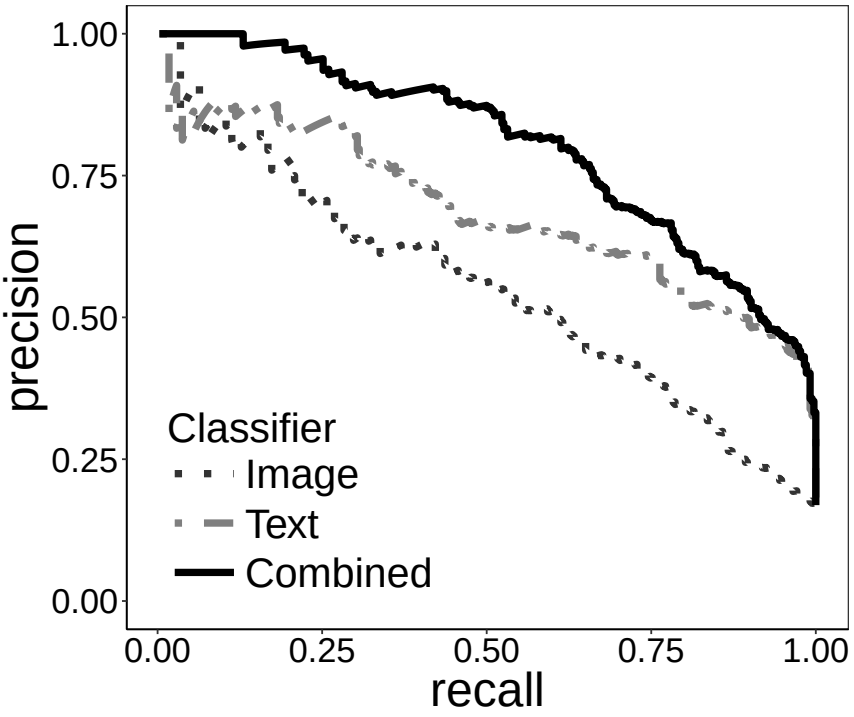
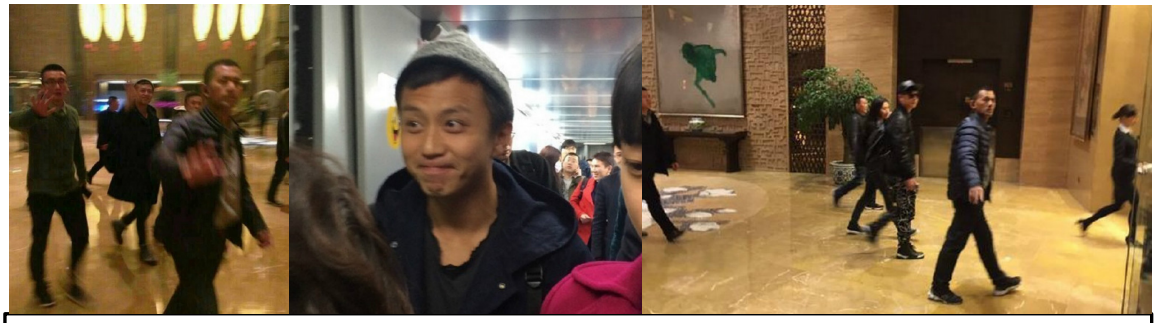


Figure 4: Precision-recall curve of the CNN image model (dotted line) text-based CNN-RNN model (dotted dash line), and the combination of the text and image classifier (solid line), in the first-stage based on out-of-sample validation.

Overall, the performance of the first-stage classifier is strong, but we wanted to do better than correctly classifying 63% of posts as collective action (precision 0.63) and correctly identifying 79% of collective action posts (recall 0.79). In particular, we want to improve precision, to make sure more posts we classify as collective action meet our definition of collective action. To do so, we need to reduce the number of false positives. We systematically examined the false positives from the first-stage classifier by taking a random sample of 2000 false positives, and examined them by hand. We found two types of false positives. The first were posts made by government-related social media accounts that describe how public grievances have been resolved. Figure 7 provides an example of a false positive, where a local government is publicizing how it was able to resolve worker grievances over owed wages. From reading the post in Figure 7, it is possible that workers did protest,

Why is this type of phenomenon often seen outside the gates of government offices? People are holding onto old ideas, should leaders all be extremely honest and noble? Where is the problem? Will such a country prosper? Will it endure?| Weiyang Middle Road 为什么市政府大门口经常会出现这种现象? 百姓迂腐，难道我们 的领导们都很清廉高尚嘛?问题到底出在哪?这样的国家到底会不会繁荣昌 盛?会不会长久下去?| 渭阳中路	In front of the Office of Letters and Visits……X ah! Dear mother! The aggrieved are kneeling for you, standing for you! Don't kneel. 国家信访局门前.....X 啊!亲娘!冤民给你下跪了站直了!不要 跪

Figure 5: Weibo posts about collective action correctly classified as related to collective action (true positives) from the first-stage classifier.



Running Man Season 2 began filming in Chengdu. Deng Chao, Angelababy, Li Chen among other actors were surrounded by a mob. Recently, some netizens posted photos of the Running Man Season 2 actors gathering in Chengdu. Deng Chao, Li Chen, Zhang Kai and Angelababy, Li Chen were all present. And Han Geng also arrived in Cheng Du around midnight. It was just revealed online that he will collaborate with Fan Bingbing on the first show of the season.
奔跑吧兄弟 2 成都开录邓超 baby 李晨等遭围堵. 近日，有网友曝 光"跑男"第二季成员们聚集成都的照片。邓超、李晨、郑恺、baby 等均现 身。而韩庚也于凌晨抵达成都，网曝他将与范冰冰共同录制第一期。

Figure 6: Weibo post containing one of the words in K, "surrounded by a mob," correctly classified as not related to collective action (true negatives) from the first-stage classifier. The images in this figure are from a set of six associated with one Weibo post.

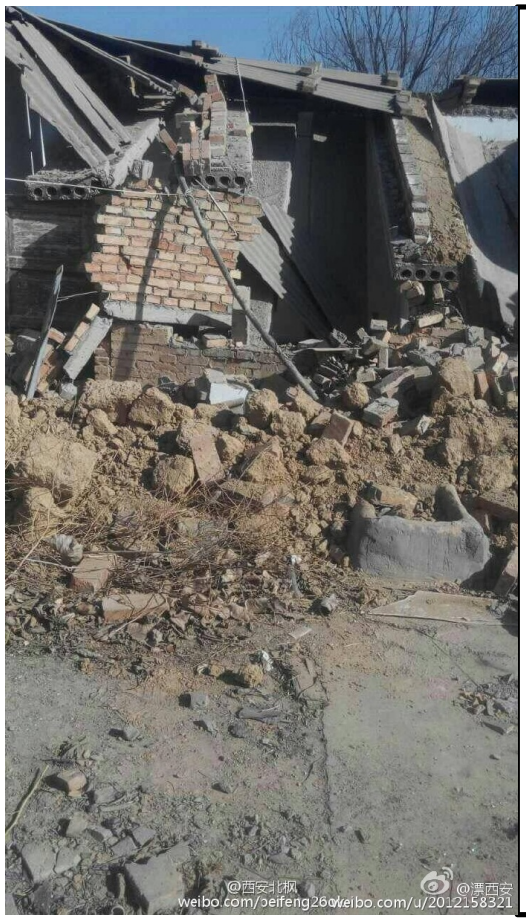
...In recent days, eight Feidong migrant workers asked the sub-district for help in obtaining back wages totaling 40 thousand yuan...the sub-district procuratorate immediately launched a legal aid program for migrant workers, and after 7 days of effort, migrant workers were paid the back wages. Look, migrant workers event sent a banner for the staff
…近日，肥东县古城镇陈天扬等8名肥东县农民工，向海棠街道寻求帮助，他们反映辖区一建筑工地工程承程包商，去年至今共拖欠工资4万余元。街道司法所立即启动为农民工讨薪法律援助程序，经过7天的工作，承包商程某终于偿还了工钱。瞧，农民工还为工作人员送来了锦旗呢

Figure 7: Weibo post made by government accounts describing their effort in addressing public complaints incorrectly classified as related to collective action (false positive) from the first-stage classifier.

and then the government responded; however, this post does not meet our definition of collective because we do not know for sure whether there was an on-the-ground protest with three or more individuals.

The second type of false positive are social media posts that describe issues, problems, grievances, which can lead to collective action, but not in the social media posts in question. For example, Figure 8 shows a Weibo post made by someone whose house was demolished. This individual was not able to reach an agreement on compensation for housing demolish-

ing with the government, so someone (presumably at the governments behest) destroyed the building in the middle of night. The image that accompanies the text shows the ruins of the building. Housing demolition appears frequently as a motive for collective action,



Because the government was not able to reach an agreement on housing demolition after negotiating multiple times, on the night of January 9, 2015 around 3am, a person of unknown identity drove a front loader into my house, knocking it down. Right now, my son is serving in the army, exerting himself for the country, the family of a service member comes to this harm
由于政府拆迁互相多次谈判都没有达成协议，2o15年1月9日夜晚3点左右有不明身份的人开着铲车在我居住人的三间房子后墙推到了我的住房，目前我的儿子正在部队服役，为国出力，军人家属遭到这样不明的侵害

Figure 8: Weibo post that does not meet our definition of collective action incorrectly classified as related to collective action (false positive) from the first-stage classifier.

but in the case of Figure 8, there is no evidence of any event that meets our definition of collective action presented at the beginning of Section 3. As a result, this post is incorrectly classified as a false positive. The two quotes below are examples of text-only social media posts incorrectly classified in the same way.

> What is the government of this country doing! Forced demolition, forced land taking, corruption, and taking bribes? 这个国家的政府到底是干什么的!是

不是强征强拆，贪污受贿?

What is law enforcement? Why arrest those who are just trying to demand their wages? The police are recklessly arresting people, beating people without trying to distinguish right from wrong! Is rightfully demanding wages a crime?什么执法过当? 人家只是讨薪干嘛抓人? 警察不问青红皂白就胡乱抓 人打人! 人家正当讨薪是罪犯吗?

These posts contain words describing issues that often motivate collective action in China—forced demolition (强拆), forced land-taking (强征), corruption (贪污), and unpaid wages (讨薪). These posts contain words that frequently appear in discussion fo collective action—police (警察), arrest (抓), beating (打人). The first-stage classifier is not effective in distinguishing between social media posts about collective from social media posts that talk about the same issues, complaints, and grievances but do not meet our definition of collective action. The main reason is that in the training data, the negative examples (posts unrelated to collective action) are much less likely to contain the words, phrases, images related to issues motivating protest than the positive examples.

To address the first type of false positive, we exclude posts from output of the first-stage that come from government or Chinese Communist Party accounts. This includes accounts of national and subnational governments, party offices, bureaucracies, as well as state media outlets. To address the second type of false positive, we use a second-stage classifier trained on data with a larger number of negative examples—posts that discuss issues and grievances such as housing demolition, police, and corruption, but which do not describe collective action events.

3.2.4 Second-Stage Classifier

The second-stage classifier uses the same CNN model as the first-stage classifier (the CNN model whose architecture and training method is described in Section 3.2.1).[39] However, we retrain the CNN-RNN model for the second stage (the architecture and training method remains the same as what is described in Section 3.2.2). We have a team of research assistants code 40,505 posts with predicted probability greater than 0.2 from the first-stage CNN-RNN model.[40] Four undergraduate and masters students who are native Chinese speakers identified by hand posts related to offline collective action.[41] Among the 40,505 posts, 9,761 pertained to offline collective action, and 30,744 did not. These 30,744 negative examples are crucial because they are much more likely to discuss issues and grievances that can motivate collective action than the negative examples used to train the models in the first-stage. To make this training data balanced, we take a sample of 20,983 (30,744−9,761) posts from the Wickedonna Dataset so there are also 30,744 positive examples. These posts are used to train a new CNN-RNN model for the second-stage classifier.

We use the CNN and CNN-RNN models to make predictions of 590,692 images and 718,243 texts coming from posts with predicted probability above 0.2 from the first-stage classifier.[42] We set a low probability threshold to maximize recall (recall at this threshold is 0.9[43]) to ensure that most positive cases from the first-stage classifier entered into the second-stage.[44]

Like in the first-stage, if a Weibo post contains both text and images, we combine the two predicted probabilities to obtain a single predicted probability for each post using Equation 2. The optimal α and β for the second-stage classifier are 0.18 and 1.04, respectively. As in the first stage, α is smaller than 1, suggesting that relatively more information is based on text classifiers, and again β is larger than 1, suggesting that a post

with both text and images indeed is slightly more likely to be about protest than those with only text. However, α is decreasing from the first- to the second-stage, suggesting that the information taken from images is relatively less important in the second-stage, which makes sense because many of the error in the first-stage were due to protest-related words appearing in the text of the posts, and we retrained the CNN-RNN model to better deal with this text data. We consider a post as being related to collective action if the combined predicted probability is greater than 0.65, which we selected to maximize $F1$ score based on out-of-sample validation.

In cross-validation, the two-stage classifier performs extremely well, with a maximum F_1 score of 0.94 (precision = 0.93, recall 0.94). For out-of-sample validation, we again use our test set of 10,000 posts (described in Section 3.2.3). The left panel of Figure 9 shows the precision-recall curve based on random guess (dot-dash line), cross-validation (dotted line), and out-of-sample validation (solid line).[45] As expected, precision and recall are better for

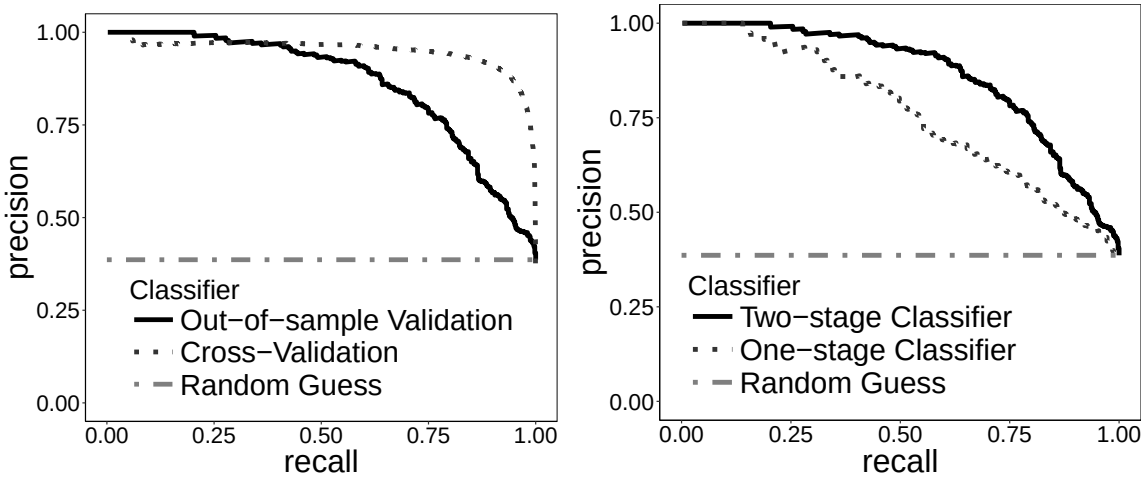


Figure 9: Precision-recall curve based on random guess (dot-dash line), cross-validation (dotted line), and out-of-sample validation (solid line) from the two-stage classifier in the left panel; precision-recall curve of the first-stage classifier (dotted line) and two-stage classifier (solid line) in the right panel.

cross-validation than for out-of-sample validation, but both are strong. The right panel of

Figure 9 compares the out-of-sample performance of the first-stage classifier alone with the performance of the output of the two-stage classifier. The out-of-sample performance of the two-stage classifier is much better than that of the first-stage alone, with a maximum F_1 score of 0.83 (precision = 0.84 & recall = 0.80). This means that at the end of the two-stage classification process, 84% of the posts CASM predicts to be about collective action posts are about collective action, and CASM captures 80% of the human-coded collective action events from our out-of-sample validation data of 10,000 posts. Note that 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.

From this two-stage classifier, we identify a total of 283,427 $T_{protest}$ posts out of 9.5 million that are likely discussing collective action between January 1, 2010 to June 30, 2017.

3.3 Identifying Collective Action Events

The final step of CASM is to identify unique collective action events from the posts identified by the two-stage classifier. We do so by adopting a rule-based approach that utilizes the temporal, spatial, and text information contained in the posts $T_{protest}$. We extract two pieces of data from each posts in $T_{protest}$. The first is the date of the post. The second is the location of the post. The date of the post is included in the metadata of every post we gather from Weibo, so this step is straightforward.

3.3.1 Identifying Location

The identification of the location of the post is less straightforward. China is divided administratively below the central level into provinces; provinces into prefectures; prefectures into counties; counties into townships; and townships into villages and neighborhoods (neighborhoods are the urban equivalent of villages). We want to locate collective action events to these administrative divisions because collective action events in China often involve the government. Even when the target of protest is not the government, protesters often ask for governmental intervention. As a result, it makes sense to align location with existing administrative boundaries to identify unique collective action events.

When a Sina Weibo user makes a post, the user has an option to share the exact location where the post is being made. Only 4.4% of posts in $T_{protest}$ (12,471) have this attribute. When this information is available, we use the longitude and latitude information from the Weibo metadata to locate the post to counties in China.

When this precise geo-location data is not available, we extract location information from the text of the post. We take a list of names of provinces, prefecture, and counties from China's National Bureau of Statistics,[46] and look for these names in the text of the post in $T_{protest}$. We can find the prefecture name from the text of the post for 75.6% (214,258) of posts in $T_{protest}$. We can find the county name from the text of the post for 40.7% (115,358) of posts in $T_{protest}$,[47] There are approximately 300 prefectures and 3,000 counties in China. The county and prefecture are important levels of administration, and often targeted by protesters because they have the authority to penalize grassroots officials for corruption and adjudicate disputes with companies and commercial interests.

When we compare the county and prefecture location identified by our text-based extraction method against the location identified through longitude and latitude, we find that our method performs well—the county or prefecture name extracted from the text of the

Weibo post matches the county or prefecture identified by longitude and latitude 95% of the time. The remaining posts may contain some indication of location, e.g., "road" (路), "avenue" (大道), "village" (村), "market" (市场), "elementary school" (小学), or it may contain province names, but we were not able to make use of these information in a consistent manner to determine location. We take a conservative approach and discard posts that we cannot geo-locate.[48] This means the number of collective action events we identify will be an under-estimate relative to the posts we identify.

3.3.2 From Location to Events

We use the location information described above and combine posts into events by day, where day is defined as a 24-hour period from 12:00am to 11:59pm China time. Specifically, when we cannot identify the county associated with a post, we consider all posts made within the same prefecture on the same day to be the same event. When the county can be identified, we consider all posts made within the same county on the same day to be the same event. Posts we can locate to a county level are considered distinct events from posts we can locate to the prefecture level, but which do not contain county information, on the same day. We group by day because few protests (less than 1%) are reported on social media for more than one day.

To illustrate this approach in practice, suppose there are five posts that reference Prefecture A on January 1st. Two of those posts do not contain county names. Among the three posts that do contain county names, two posts reference County X, and one post references County Y. According to our grouping method, there are three collective action events on January 1st: one event in Prefecture A (described in two posts), one in County X (described in two posts), and one in County Y (described in one post).

There are a number of shortcomings to this method. We may be grouping separate events into one event if there is more than one collective action event occurring on the same

day in a prefecture or county. We could also be inflating the number of events. Using the above example—the two posts referencing Prefecture A without mentioning any county names could reference the collective action event occurring in County X, and we would mistakenly count these as two separate events.[49] We would miss cross-regional protests, which are rare in China but would be of substantive interest.

From the 283,427 posts about collective action in $T_{protest}$, we can identify location for 226,729 posts, from which we identify 142,427 unique collective action events. Going forward, we refer to this dataset of 142,427 unique events as the CASM-China dataset. This means on that on average, each collective action event is discussed in 1.59 posts. This suggests that CASM is able to recover collective action events that receive limited overall attention on social media.[50]

In future iterations of CASM, we hope to explore alternative methods of event grouping. We could continue with location-based grouping, but work to make use of additional location information, such as well-known locations that are not administrative regions (e.g., Tiananmen Square, Beijing Railway Station, Zhejiang University). We could also experiment with grouping based on issue in addition to location and time.[51]

4 CASM Ouput and External Validity

The 142,427 collective action events that constitute the CASM-China Dataset occur in regions throughout China. Figure 10 shows the logged count of CASM-China events by prefecture. Darker colors correspond with more collective action events; lighter colors fewer events, and prefectures in gray are those for which we did not identify any collective action events. The regions where we do not identify collective action events over the seven and a half year period are clustered in ethnic minority regions such as prefectures in Tibet, Xinjiang, and Sichuan, or in military-controlled areas such as counties in Hainan. The

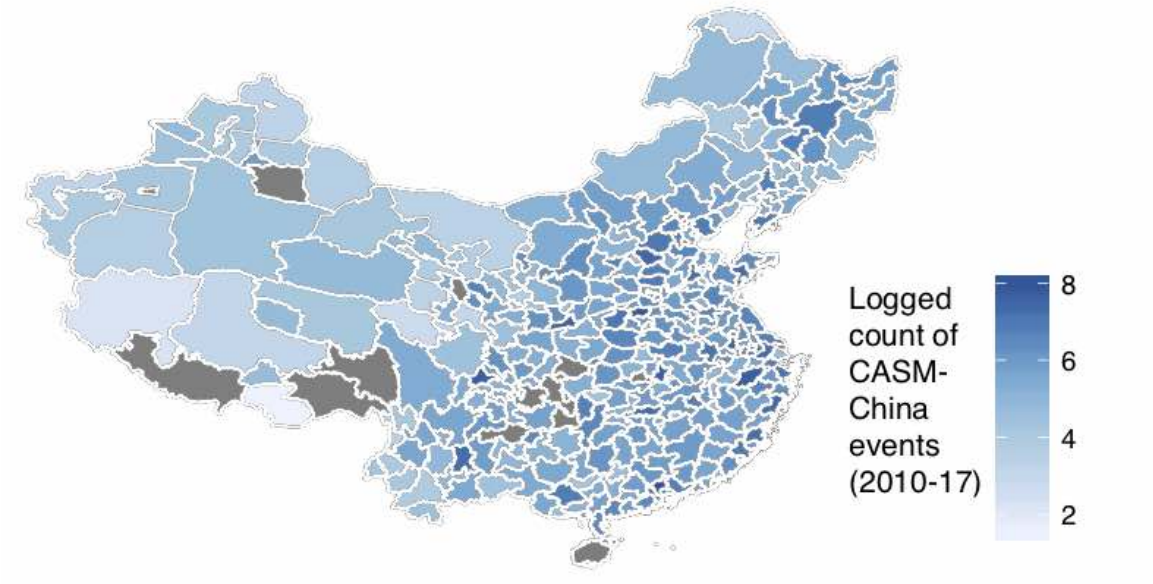


Figure 10: Log count of CASM-China events by prefecture (January 2010 to June 2017)

lack of data from Tibet and Xinjiang may reflect the imposition of more stringent forms of repression and internet controls in these regions by the Chinese government, and lower usage of Chinese-language social media platforms.

The solid black line in Figure 11 shows the monthly count of events in CASM-China from January 1, 2010 to June 30, 2017. The number of events increases from 2010 to 2013, and slowly declines after. The 2010 to 2013 increase is likely due to the growing popularity of Weibo and increasing availability of data. The 2013 to 2017 decline likely in part reflects the declining popularity of Sina Weibo. To account for the change in the popularity of the Weibo platform, we gather posts containing a Chinese idiom we do not expect to relate to collective action.[52] Usage of this idiom on Weibo also declines from 2013 to 2017 (see Supplemental Appendix). If we use this idiom as an indicator of Weibo's declining popularity and divide the count of CASM-China posts by the count of posts containing this idiom to control for declining usage of Sina Weibo, we find, in Figure 11, that volume of collective action events remains steady overall from 2013 to 2017 but experiences short term fluctuations—a spike in the relative number of events identified by CASM at the

beginning of 2015 (January and February) and near the middle of 2017 (May and June). This overall result goes against the prevailing perception that collective action is steadily

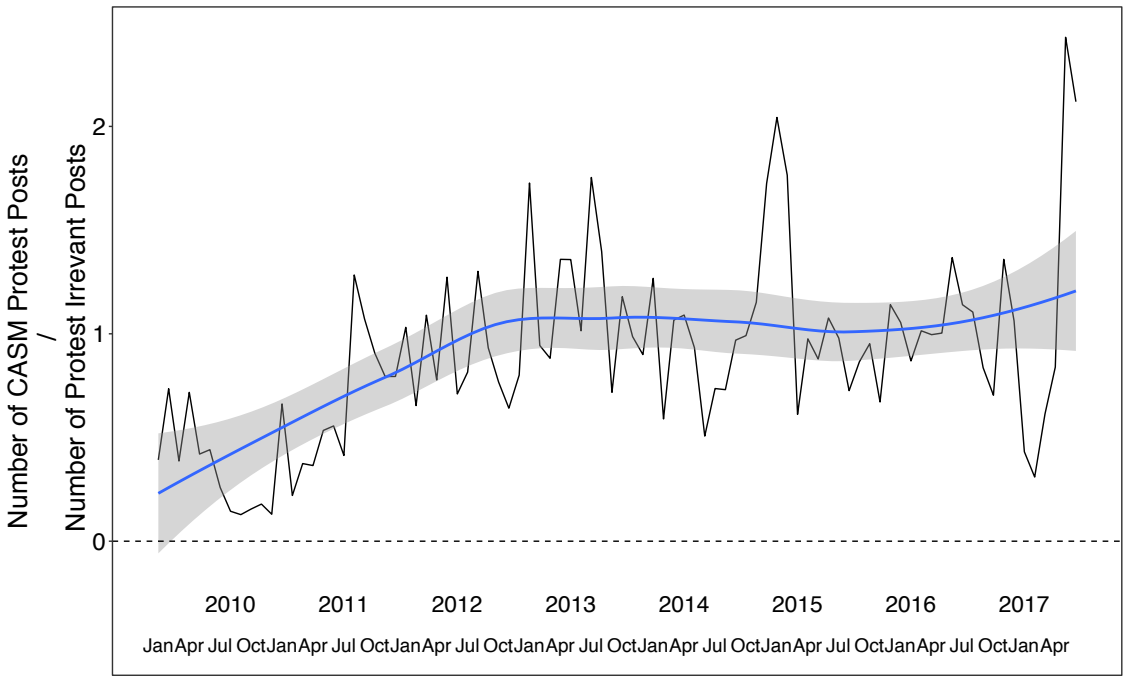


Figure 11: Number of CASM-China posts divided by the number of posts that contain an idiom unrelated to collective action, with loess smoothing.

increasing in China. Although we do not know the reason for this time trend, it does overlap with the Xi Jinping era, which has been characterized by more stringent social and political controls.

The focus of this paper is to describe how CASM works, and to provide the general contours of CASM-China. However, we recognize that the text and images of $T_{protest}$ contain much more information about collective action events than simply the time and location of occurrence. Here, we provide a first-pass look at two features of collective action events—the form of protest and the issues motivating protest, which we identify by using keywords generated from close reading of posts in $T_{protest}$ and from existing research

on collective action in China (Cai, 2010a; Chen, 2011; Lee, 2007; O'Brien and Li, 2006; Perry, 2008; Qin et al., 2017; Weiss, 2014) (see Supplementary Appendix for the set of keywords).[53]

Following Almeida (2003), we categorize collective action events into three main forms. The first form is what we call "conventional" collective action events, such as street marches, strikes, public gatherings, public demonstrations, public group petitions. The second form is what we call "disruptive" collective action events—for example, occupation of buildings, occupation of land, construction of barricades, and cutting off power supplies. The third form is what we call "violent" collective action events, including armed attacks and physical conflicts with government officials. An event is considered to be of a particular form if posts pertaining to the event contain one or more of the keywords in that category (see Supplementary Appendix for the set of keywords). An event is placed in the "violent" category if it contains any of the keywords in this category, even if it contains keywords belonging to either of the two other categories. An event is placed in the "disruptive" category if it contains keywords in this category and the "conventional" category. We code posts in this way because violent and disruptive forms of collective action, which incur higher costs, are of particular substantive interest (Lorentzen, 2013). We find that 39% of collective action events in CASM-China are conventional in form, 37% exhibit disruptive characteristics, and the remaining 24% exhibit violent characteristics.

We also examine the issues that motivate protest using hand-curated keywords (see Supplementary Appendix for keyword list). We focus on 11 types of issues, which China scholars have identified as important motives for collective action in China today (Cai, 2010a; Dimitrov and Zhang, 2017; Goebel, 2017). These issues include (in alphabetical order):

1. Education: protests by parents over the difficulty of enrolling their children in public

schools, over inequalities in educational access based on wealth and geography, and over the perceived bias and corruption of school administrators and teachers.[54]

2. Ethnic / religious: collective action by ethnic minorities such as the Uyghurs in western China as well as religiously motivated collective action such as those occurring in the aftermath of Christian church demolitions in eastern China.[55]

3. Environment: collective action over environmental issues such as air pollution and the construction of chemical plants (Deng and Yang, 2013).

4. Fraud / scams: protest over fraud, scams and the lack of consumer protectors, such as those that erupted following losses sustained in risky peer-to-peer lending platforms.[56]

5. Homeowner / property: collective action motivated by conflicts over property ownership, primarily related to corruption by real-estate developers and property management companies.[57]

6. Medical: protest over medical disputes—for example, family members protesting against hospitals for negligence and malpractice (Liebman, 2013).

7. Pension / welfare: collective action over welfare provision, especially pensions (Hurst, 2004; Hurst and O'Brien, 2002).

8. Rural / land: Collective action due to forced land taking, and other land-related conflicts in rural areas (Guo, 2001).

9. Taxi: protests by taxi drivers, which have intensified in recent years over fees imposed by local governments as well as competition from ride-sharing companies.[58]

10. Unpaid wages: collective action due to unpaid wages by workers and migrant workers (Blecher, 2002; Su and He, 2010).

11. Veterans: protests by veterans over welfare and benefits (Diamant, 2010; Tong and Lei, 2010).

Although certain of these issue categories are sub-categories of larger issues—e.g., taxi driver protests and protests over unpaid wages are all labor issues—we include more specific categories because they have been of interest to China scholars. Instead of mutually exclusive categories, if an event contains posts with keywords across issues, we place the event in multiple categories, and we reweigh the distribution so that the category proportion of each issue sums up to one.[59]

Among events containing posts with keywords related to the issues described above,[60] just over a quarter of CASM-China events relate to unpaid wages (26%), slightly less than a quarter relate to conflicts over property (24%), and slightly less than a quarter related to conflicts over land (23%). The remaining 30% or so of events fall into the remaining eight issue categories (for details, see Table 3).

4.1 Comparison with Other Protests Datasets in China

We compare CASM-China against other datasets of collective action. This is important because it makes the biases and limitations of CASM and of the resulting data clearer, so that any analysis conducted on these data can be interpreted more appropriately.

We use three datasets of collective action based on newspaper data: the Global Database of Events, Language, and Tone (GDELT), the Integrated Conflict Early Warning System (ICEWS), and WiseNews.[61] GDELT takes an unsupervised machine learning approach to identify events of interest, including collective action events, from global news sources.[62] ICEWS also monitors global news agencies to detect political events, with an emphasis on accuracy. WiseNews is a dataset of collective action events we generate by applying CASM on a corpus of more than 1500 major Chinese, Hong Kong, and Taiwan newspapers from

the WiseNews Database (Shao, 2017). In addition to newspapers, the WiseNews Database also contains social media data from WeChat, another social networking site in China.[63] Details of these three comparison datasets can be found in the Supplementary Appendix.

We use 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 (CLB), which documents labor-related protests.[64] Both datasets have been used by scholars of China to study collective action (Dimitrov and Zhang, 2017; Goebel, 2017).[65]

Because these datasets cover different time periods, we compare these datasets for a six month period from January 1, 2016 to June 30, 2016. Table 1 shows the number of collective action events identified by CASM-China and all of the comparison datasets. Table 1 shows that CASM-China identified 10,432 events during the first half of 2016. The Wickedonna Dataset contains 11,085 events, CLB 1,455 events, GDELT 299 events, WiseNews 276 events, and ICEWS 25 events during the same period. The low number of events identified by GDELT and ICEWS likely reflects limitations on reporting placed by the Chinese government on foreign media. The low number of events in WiseNews may be driven by our method of identifying collective action events (by applying CASM on newspaper data); however, it likely also reflects Chinese government constraints on media reporting of

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	10,432	
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	65%
China Labor Bulletin	Mixed	2011-	1,455	75%

collective action given the difference is in orders of magnitude.

Also shown in Table 1, slightly over half of the collective action events identified by GDELT and ICEWS are found in CASM-China—56% of events in GDELT, 52% of events in ICEWS. Even though a relatively small number of protests are reported by international news outlets, CASM-China has low coverage of these collective action events. This is primarily due to the emphasis of foreign media on ethnic and religious conflict, which appears relatively rarely on social media. Eight-eight percent of collective action events in WiseNews are in CASM-China. Among protests reported in the China Labor Bulletin, 75% of events are covered by CASM-China, and 65% of events in the Wickedonna Dataset are covered by CASM-China. Because data for CLB is based on the subset of data from the Wickedonna Dataset, especially during the first half of 2016, we examine in greater depth the 35% of collective action events identified by the Wickedonna Dataset that are not in CASM. We find that 15.8% of the events in the Wickedonna Dataset are not detected by CASM-China because they do not contain any keyword from our dictionary K; 10.4% are not identified because the posts are no longer found on Sina Weibo, likely due to censorship; and the remaining 8.3% are not found likely due to Weibo's restriction on data collection.

Table 2 shows the proportion of events in CASM-China and the comparison datasets that are conventional, disruptive, and violent. GDELT events contain the largest proportion

Table 2: Comparison of the Form of Collective Action (Jan. 1 to Jun. 30, 2016)

	CASM	Wickedonna	CLB	GDELT	WiseNews
Conventional	39%	46%	36%	26%	28%
Disruptive	37%	30%	52%	44%	56%
Violent	24%	23%	13%	30%	17%

of violent events (30%), followed by CASM-China (24%) and Wickedonna (23%). The presence of violent collective action in GDELT contains aligns with existing research on biases in media reporting toward events that are more sensational in nature (when media

is not state-controlled). Disruptive events appear with highest prevalence in WiseNews (56%) and CLB (52%), followed by GDELT (44%), CASM-China (37%) and Wickedonna (30%). This result is interesting because we are applying CASM tuned on Weibo data on WiseNews data to identify collective action events, which might make the distribution of events in WiseNews more closely resemble that of CASM (more conventional events), and what we know about media reporting should lead us to see more reports of violent protests, but we see neither of these outcomes. A majority of collective action events reported in WiseNews are disruptive. One possible explanation for this is that the Chinese government may prohibit Chinese media outlets from emphasizing violent protest, and in an attempt to capture audience, Chinese media outlets focus relatively more on disruptive events. More research would be need to test this hypothesis. Finally, Wickedonna contains the largest share of conventional events (46%), followed by CASM-China at 39%.

We compare the distribution of issues in CASM-China to that of other datasets by applying the same keyword approach to data for the Chinese language sources. Because the GDELT data is in English, we place them into these issue categories by hand. Table 3 shows the proportion of events containing keywords in each category, with rows in descending order based on the proportion of events related to each issue for CASM-China. We find that CASM-China may identify relatively more collective action events related to rural land disputes (23%)) than other datasets (e.g., 12% in Wickedonna, 6% in WiseNews, 4% in GDELT). This is striking because we might expect rural usage of social media to lag behind that of urban areas, biasing events in CASM-China away from rural events. However, rural social media usage has expanded dramatically in the past decade (McDonald, 2016), and we see that reflected in relative share of rural and land related collective action events captured by CASM-China. As expected, CASM-China identifies relatively far fewer events related to ethnic and religious conflict (0.49%) than GDELT (40%). This reinforces the fact that CASM-China identifies relatively fewer collective action events from minority regions

Table 3: Comparison of Issues Motivating Collective Action (Jan. 1 to Jun. 30, 2016)

	CASM	Wickedonna	CLB	GDELT	WiseNews
Unpaid wages	26%	27%	73%	19%	17%
Homeowner / property conflicts	24%	27%	12%	0%	61%
Rural / land conflicts	23%	12%	0%	4%	6%
Educational dispute	8%	10%	2%	8%	0%
Medical dispute	7%	8%	0.18%	0%	0%
Taxi	4%	5%	8%	0%	0%
Environmental	3%	4%	0.38%	15%	6%
Fraud / scams	3%	4%	0.85%	7%	11%
Pension / welfare	1%	2%	2%	0%	0%
Ethnic / religious	0.49%	0.54%	1.76%	40%	0%
Veterans	0.44%	0.45%	0%	0%	0%

of China, while international media focuses on ethnic tensions.The distribution of issues in CLB is heavily skewed toward issues related to unpaid wages, which is expected because the focus of CLB is on labor. The distribution of events identified in WiseNews heavily emphasizes conflicts over property and homeownership, followed by unpaid wages, and by fraud and scams. The WiseNews results are again interesting because our method of identifying collective action in WiseNews should bias the type of events found in WiseNews toward those found on social media (unpaid wages, property conflict, land conflicts), but what we see is that there is very little Chinese news reporting on land conflicts compared to social media reports, and no media reporting on several categories of issues, such as ethnic and religious issues. It is possible that we do not find collective action events related to those issues because Chinese media is censored on these topics or because Chinese media talks about these topics in ways that are different from that of Chinese people.

Because CASM-China shares many similarities with the Wickedonna Dataset, we also compare them over time. Figure 12 shows the count of events from the two datasets from 2010 to 2017. The post-2013 trend between the two datasets looks very different. Despite a steady increase in the number of events in the Wickedonna Dataset, while there is a decrease

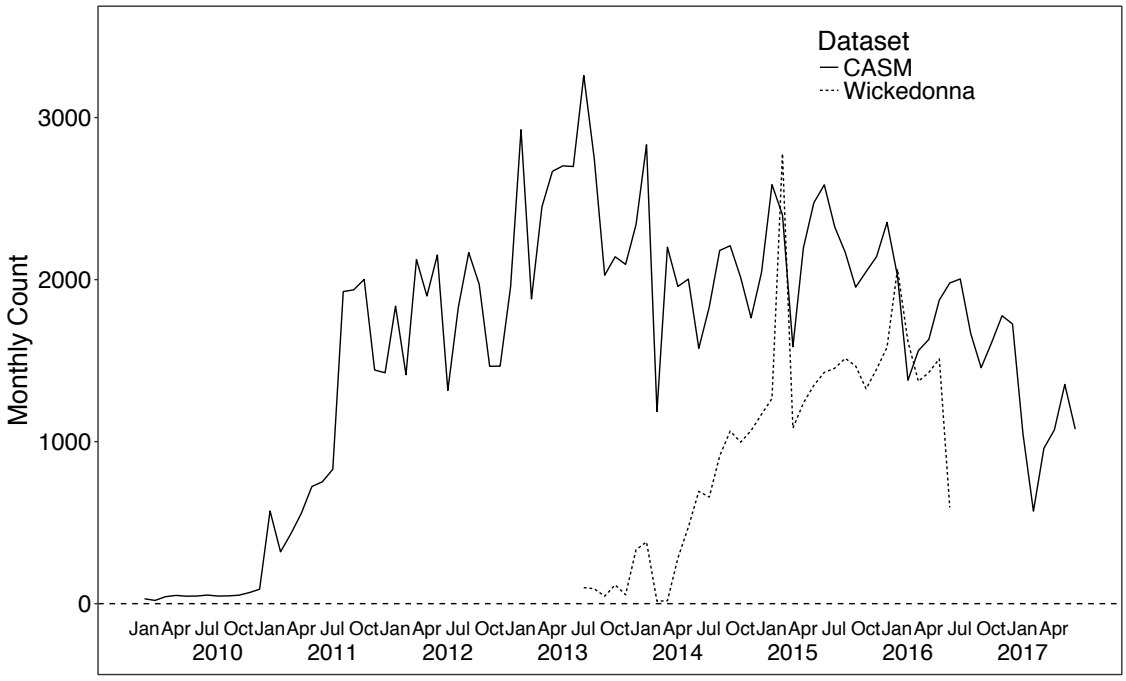


Figure 12: Monthly count of CASM-China collective action events (solid black line) compared with monthly count of events from the Wickedonna dataset (dotted line).

in the number of CASM-China events. We know from previous analyses in Figure 11 that some of the decrease in CASM-China is due to the declining popularity of Sina Weibo, but even after controlling for usage of Weibo, the time trend in the number of protests between CASM-China (stable from 2013-2017) and Wickedonna (increasing from 2013 to 2016) are different. Three possibilities account for this difference. The first is that the activists who hand-curated the Wickedonna Dataset improved their ability to identify collective action events over time. The second is that the definition of protest in the Wickedonna Dataset expanded overtime. Third, discussions of collective action moved to other social media platforms, which are captured by Wickedonna but not CASM-China, at rates greater than the general decline in popularity of Weibo.

These comparisons show that no dataset of collective action, including CASM-China,

should be considered "complete." The output of CASM-China is biased but can complement events identified in hand-curated datasets, international media reporting, and Chinese media reports.

4.2 Online Censorship

Another potential source of bias is online censorship. As discussed in Section 2, censorship can limit the use of social media as target data for identifying collective action events by generating self-censorship and by limiting the diffusion of knowledge around protest. In addition to these general effects of censorship, in China there is the additional problem that online censorship is specifically aimed at removing discussions of collective action (King et al., 2013, 2014). How can social media data be used to detect collective action events in China when this is the case? The answer to this seeming contradiction lies in the recognition that content removal in China is post-hoc, focused on bursty (viral) online discussions, and incomplete. Censorship of collective action is not based on keywords (King et al., 2014). 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, a few posts often escape censorship.

We empirically assess the impact of Chinese censorship. We use a corpus of Weibo data collected in real time in January 2018, before the Chinese government can remove the subset they deem objectionable. Among these posts, we find 113,081 posts containing at least one of the 50 keywords in K. We apply CASM to these 113,081 posts. We identify 7,489 posts related to real-world collective action, and we can geo-locate 4,979 to the

prefecture or county level. From these 4,979 posts, we can identify 937 unique events. Among these posts, only 267 (5.4%) were later censored, leading to a loss in identifying 25 unique collective action events (2.7%)

This analysis confirms our expectation that post-hoc content removal does not have a large influence on the ability of CASM to detect collective action events that have been reported online in China. However, this does not account for other forms of online censorship such as self-censorship—individuals unwilling to post on social media for fear of reprisals—and other effects of censorship—limited on diffusion of knowledge of collective action online.

5 CASM Beyond China

The framework and approach of CASM can be applied to other regions of the world, to other linguistic, cultural, and political contexts. The following aspects of CASM are generally applicable: using keywords to first select social media posts to improve precision because of rarity of online discussions about collective action events; using CNN models to classify image data and CNN-RNN models to classify text data; using a second-stage classifier to differentiate discussions of offline collective action events from social media posts with similar terminology but do not related to collective action events; and using the temporal and spatial information of social media posts to identify unique events.

However, CASM tuned on Weibo data cannot be applied wholesale to non-Chinese language social media data. In order to adapt CASM for other languages and countries, we need to consider 1) where / how to collect data, 2) data to identify keywords K and train text / image classifiers, 3) whether to include the second-stage classifier, and 4) how to combine posts into events.

Data: Outside of China, the main source of social media data is likely to be Twitter, which has widespread, global adoption, and whose data is relatively accessible for researchers (Steinert-Threlkeld, 2018). This means this that CASM may not be applicable in countries like Myanmar, where social media is dominated by Facebook, and where discussions of collective action take place on Facebook, because Facebook data is much less accessible for academics.[66]

Identifying keywords: To identify keywords and to train the first-stage deep learning algorithms, we relied on a large set of data on collective action curated by two Chinese activists. Equivalent training data will not always be available in other contexts, and the unavailability of training data will limit the application of CASM. However, there is a growing literature that has relied on large-scale social media data to study collective action events around the world,[67] and there is emerging research analyzing social media imagery of protest.[68] Although studies are not focused on using social media data to identify protest, they nonetheless represent collections of social media data related to protest that can be used to identify keywords (K) and can be used as training data for the first-stage classifier. In terms of retraining the deep learning algorithms, the text-based CNN-RNN model must be retrained if applied to non-Chinese social media data. However, the CNN model for image classification may be more easily transferred outside of China and used to identify protests elsewhere because it is not language dependent. Researchers in other regions could take our pre-trained CNN model as is, or only train one or two of the final convolutional layers if their data is limited.

Second-stage classifier: In contexts outside of China, we still expect that a second-stage classifier will improve performance. Social media is user generated, and it is used for claim-making (Koopmans and Statham, 1999). The motives for offline collective action

may overlap with other claims made on social media, and a second-stage classifier would help differentiate between these different content; however, a second-stage classifier for other country contexts will require human coding on the output of the first-stage classifier. Thus, whether to include the second-stage classifier is a decision that a researcher would need to make based on the performance of the first-stage classifier relative to the cost of human coding.

Grouping posts into events: Finally, how posts are combined into events will differ in different contexts because the availability of geo-located metadata, the type of geographical units used, and the difficulty of extracting location from text will vary by country and by language. For example, in democracies like the US it may make less sense sense to use government administrative regions as the base location unit because many of the collective action events involving government in China are precipitated by grievances that can be addressed by the legal system in countries with rule of law. Advances are being made in the extraction of geo-location data from text (Lee et al., 2018), which we expect will aid in this process.

In extending and using CASM beyond our application here, several other consideration merit discussion. Social media changes quickly over time in terms of topics of discussion, platform features and algorithms, as well as the emergence of new apps and technologies. The CNN and CNN-RNN models in CASM will need to be retrained repeatedly over time to capture changes in how users communicate on social media. Care needs to be taken when making comparisons over time, and needs to account for changes in the popularity of social media sites (as we have done with Weibo in Figure 11). For example, if the framework for CASM is applied to a different country on Twitter data, we need to consider the rate of Twitter penetration in that country when examining changes in the number of collective action events over time. In addition, censorship will vary across countries. In

China, internet content providers censor content quickly and thoroughly in accordance with government demands, resulting in bursts of censorship around discussions of collective action. In other countries, the market for social media is dominated by US firms that acquiesce to censorship demands slowly or impartially (Pan, 2017), such that censorship is not aimed at removing online discussion of collective action. Instead, physical repression might be used to motivate self-censorship, or internet blackout and website blocks are implemented to prevent access to information. This means the bias induced by censorship in other contexts will relate to the extent to which social media is used to discuss collective action, rather than the extent to which researchers can collect social media posts about collective action before they are removed.

6 Conclusion

This paper introduces CASM, an approach to identifying collective action events using social media data. We discuss the advantages and limitations 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 deep learning, image as data, and two-stage classification. We extensively assess the internal performance of our system through cross-validation and out-of-sample validation. We assess the external validity of the output of CASM by comparing it to other protest event datasets and by evaluating the impact of censorship. We hope these assessments show more generally how internal and external validation can help make the application of computer science methods to social science domains more usable. The implementation of CASM in China, using Sina Weibo data, results in a large dataset of collective action events with high spatiotemporal resolution spanning a seven-year period.

There are ethical considerations related to creating a system to identify collective action

events from social media data. Social media data is generated by individuals, and can contain personally identifiable information. Collective action is often a form of participation that non-democratic governments deem objectionable. Creating a system to identify these activities could face a "dual-use dilemma" in that a system created for research purposes could be used by other actors in potentially harmful ways (Miller and Selgelid, 2007; Selgelid, 2013). We describe the methods of this system here because the underlying models we use (e.g., LSTM) and/or the methods behind them are already publicly available, and because we are measuring collective action retrospectively. We believe in the need for replicable and transparent research outweights dual-use concerns in this case.

Altogether, social media data provides unique benefits as a source of data for detecting collective action events in authoritarian regimes because it provides information when other sources such as traditional 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.

ENDNOTES

[1]See http://ronfran.faculty.ku.edu/data/index.html; see Earl et al. (2004); Hutter (2014b); Rucht et al. (1999) for more complete reviews of newspaper-based protest datasets.

[2]See https://www.internetworldstats.com/stats.htm and https://pewrsr.ch/2Kct7Qu (Accessed November 5, 2018).

[3]See https://bit.ly/2Qki9aP (Accessed November 5, 2018).

[4]See https://bit.ly/1j3wBem, http://www.stat.go.jp/english/data/nenkan/1431-02.htm, https://www.stats.gov.sa/en/5305 (Accessed November 5, 2018).

[5]We also chose three because, events with three or more people is rumored to match Chinese government definition of "mass incident."

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

[7]Weibo is functionally similar to Twitter, which is not easily accessible from China, but one difference between Weibo and Twitter is that Weibo allows users to comment on a post without retweeting (similar to comments on Facebook).

[8]Based on human coding of a random sample of 10,000 posts containing keywords in K, only 7% of posts meet our definition of real-world collective action. See Section 3.2.3 for additional discussion of this sample of 10,000.

[9]See https://newsworthknowingcn.blogspot.com, which is one of the places where their data is hosted.

[10]Specifically, 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. In addition, 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.

[11]The stopword list we use can be found at https://bit.ly/2zq4sk3.

[12]We use Weibo automated searches to collect this data. Weibo returns at most 1000 posts per search so we submit search requests with extremely narrow time ranges to maximize search results. Weibo appears to limit automated searches for certain words like "march" (游行), "strikes" (罢工), and "government" (政府). This is not to say searching generates no results for these terms, but a reduced volume of results is associated

with a few of the 50 keywords.

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

[14]See https://github.com/fxsjy/jieba for more details.

[15]We do not remove stopwords when creating the word vectors in the first, embedding layer of the deep learning model used to analyze text because stopwords can provide context for other words (Dhingra et al., 2017). However, we do remove stopwords for our input into the deep learning models because it improves performance.

[16]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 discuss the GDELT system in the Supplementary Appendix.

[17]Supervised methods have been shown to outperform rule-based methods in identifying collective action events based on newspaper articles (Ramakrishnan et al., 2014).

[18]We refer readers interested in delving deeper into these methods to LeCun et al. (2015) and Bengio et al. (2015).

[19]The layer is called "fully connected" because every unit in the previous layer is connected to every unit on the next layer.

[20]VGG is the abbreviation of the Visual Geometry Group, based at Oxford University, that developed the architecture. There are many other alternative architectures such as LeNet (LeCun et al., 1989), AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), and ResNet (He et al., 2016b). On the ImageSet Classification Challenge, which is the standard evaluation criteria in computer vision research, VGGNet outperforms LeNet and the AlexNet in classification accuracy, but is outperformed by GoogLeNet and ResNet. However, we choose VGGNet because it is simple conceptually, straightforward to implement, and has many pre-trained models that perform well for applications in a wide variety of domains (Rattani and Derakhshani, 2017).

[21] For our model, the number of feature maps inside a layer are, in order, 64, 64, 128, 128, 256, 256, 256, 256, 512, 512, 512, 512, 512, 512, 512, 512. The number of feature maps increases as features become more complex.

[22]We use the Python package Keras (Chollet et al., 2015; Abadi et al., 2015), a framework to design and adapt existing deep learning algorithms as well as to implement them. We used GPUs on Amazon EC2

instances to train our models.

[23]The original model used three fully connected layers, but it overfits in our case because our goal is to classify images as representing offline collective action or not, rather than a multi-class classification task.

[24]There is a debate in machine learning about what method is best for gradient-based optimization (Wilson et al., 2017). Some argue that adaptive methods underperform stochastic gradient descent (SGD). We compared the performance of our models using adaptive gradient-based optimization and SGD. We use adaptive methods because we find that they outperform SGD for our data, even when we widely vary the learning rate of SGD.

[25]We trained our own embeddings because most pre-trained word embeddings are in English. The 20 million posts we used for training include the 9.5 million posts T_K which contain a protest-related keyword K as well as 10.5 million posts randomly sampled from geo-located posts made to Weibo in 2016. The total vocabulary size was 332,826, and the training was done on the 50,000 most frequently occurring words in this vocabulary. We also tried using word vectors obtained by using the entire Chinese language Wikipedia (zh.wikipedia.org) as the training data, but performance was not as strong. This may be influenced by the dominance of traditional Chinese characters on Chinese language Wikipedia, and government censorship of Chinese language Wikipedia. Since May 2015, Chinese language Wikipedia has been blocked in its entirety in China, and prior to 2015, pages dealing with sensitive topics such as protest were individually blocked.

[26]There is no pre-determined rule on how many layers should be used. Wang et al. (2016b) uses three convolutional layers; Sainath et al. (2015) uses two; Zhou et al. (2015) uses one; and Xiao and Cho (2016) compares the performance of two to five layers and finds three to four layers to be the most effective. Our architecture uses eight layers in the first stage and four layers in the second stage because we see no improvement by increasing layers beyond this point.

[27]We use a filter size of five, which is common in the use of CNN for natural language processing to capture semantic and syntactic relationships (Kim, 2014; Kalchbrenner et al., 2014). We use feature maps of 16, 32, 64, and 128 going from the input layer to deeper layers. The input layer has a feature map of 16, instead of a higher number, because we apply the classifier on a relatively homogeneous set of text that at contained at least one protest-related keyword. Doubling the number of feature maps is common in CNN models for both image and text analysis (Simonyan and Zisserman, 2015; He et al., 2016b; Conneau et al., 2016). The intuition is that deeper layers learn more concrete features (e.g., slogan, key phrases), which requires more feature maps.

[28]The LSTM layer has dimension 128, which is the same as the size of the last convolutional layer.

[29]This number is smaller than the total number of images, 233,288, from the Wickedonna dataset because we exclude videos and composite images, where one JPEG file contains multiple images pasted together, and we rounded down to an even number.

[30]We collected all geo-located posts from Weibo using the now-defunct geo-location API for the first half of 2016. These 261,516 images are randomly sampled from this set of geo-located Weibo posts.

[31]We collected all geo-located posts from Weibo using the now defunct geo-location API for the first half of 2016. These 200,000 posts are randomly sampled from this set of geo-located Weibo posts.

[32]Here, likelihood is assigned by a SVM classifier trained on the positive training dataset and first negative training dataset. We use SVM because we are not concerned about prediction accuracy; we simply want to identify the posts most unlikely to be about collective action. We rank the predicted probability of posts, and selected posts with probabilities in the lowest 5% quantile.

[33]See Section 3.2.3 for additional discussion of this sample of 10,000, which is our test data.

[34]We use five-fold cross validation. We calculate α and β within a 500 by 500 grid at the $(0.1, 10)$ by $(1, 10)$ region. We record the α and β that maximize the area under the ROC curve each round of the cross-validation. Then we repeat this process five times, and the final α and β are the averages of the optimal values for each round.

[35]High precision indicates that predictions minimize false positives. High recall indicates predictions recover most relevant posts about collective action and minimize false negatives. F_1 score ($F_1 = 2 * \frac{\text{precision + recall}}{\text{precision * recall}}$) is a common measure of the overall performance of the system.

[36]We had one masters student and one undergraduate student code this data.

[37]For Figure 4, we exclude the posts from the out-of-sample validation dataset where we cannot extract geo-location (see Section 3.3). If we include all 10,000 posts from the out-of-sample validation dataset, the combined classifier still outperforms the text classifier, which outperforms the image classifier.

[38]The precision-recall curve for image data (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 contained images.

[39]We do not retrain the CNN model for image classification because although we know that we communicate differently in text when describing collective action events, we do not know whether images posted to social media are different when someone is talking about an issue as opposed to talking about that issue in the context of collective action. In other words, we have no *a priori* expectation that retraining the CNN model will lead to better performance. This is borne out in practice. When we train the second-stage CNN model using images from the 40,505 posts, the performance of the CNN model is worse in cross-validation.

This suggests that there are not consistent differences in the images people post when discussing collective action vs. similar issues and grievances which are not associated with collective action events that fulfill our definition.

[40]These posts are generated from an earlier version of the first-stage CNN-RNN model, where not all hyperparameters have been tuned. This is because human coding is very time intensive and we did not want to delay human coding as we were fine tuning the CNN-RNN model since the goal of human coding was simply to identify more negative examples, of posts containing collective action related words but describing collective action events.

[41]For the coding rules they followed, please see the Supplementary Appendix; intercoder reliability, calculated with Fleiss' Kappa based on 4000 coded posts, was 0.7 among the four research assistants.

[42]The 718,243 posts already exclude posts made by government and Chinese Communist party accounts.

[43]Among these posts, less than 0.2% are false negatives.

[44]Note that here we are not maximizing the F_1 score, which is why recall is higher than what is described when maximizing F_1 in Section 3.2.3, but both precision and F_1 are lower.

[45]The data in both panels of Figure 9 exclude the posts from the out-of-sample validation dataset where we cannot extract geo-location (see Section 3.3). If we include all 10,000 posts from the out-of-sample validation dataset, none of these results are substantially different.

[46]See http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2016/index.html (Accessed November 1, 2017).

[47]We consider the districts in municipalities (Beijing, Shanghai, Tianjin, Chongqing) to be counties for this analysis. Even though municipal districts are equivalent administratively to prefectures in regular provinces, individuals living in municipalities writing about protest often mention the district name (e.g., Chaoyang district in Beijing) and rarely reference the name of the sub-district (e.g., Jianwan sub-district in Chaoyang).

[48]Another option is to count every post we cannot geo-locate as a unique event.

[49]We also considered grouping only by prefecture or only by county, but that discards a large number of posts which do not contain prefecture names and county names, respectively.

[50]This limited social media attention could be due to censorship or due to lack of interest (see Section 4.2).

[51]To systematically validate and compare these methods would require large-scale human coding of posts from the second-stage classifier into events.

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

[53]We recognize there are methodological shortcomings in this keyword-based approach, and extracting additional protest characteristics in a rigorous manner is a priority for future research.

[54]See https://reut.rs/2DV2Mmw (Accessed November 26, 2018).

[55]See https://n.pr/2OsSf3b and https://bit.ly/2FR6HTs (Accessed November 26, 2018).

[56]See https://bit.ly/2E0t7jp (Accessed November 26, 2018).

[57]See https://bit.ly/2EyGFDJ (Accessed November 26, 2018).

[58]See https://bit.ly/2P3KJMi (Accessed November 26, 2018).

[59]For example, suppose we have 10 events, where each event is described in one post. If five of the events contain keywords related to rural/land conflicts, and the remaining five contain keywords related to rural/land conflicts as well as environmental issues, the reweighted distribution of issues would be two-thirds rural/land conflicts, and one-third environmental issues.

[60]Slightly over 24% of events did not contain any of the keywords we generated. This does not mean these events are unrelated to the issues we have outlined—they could, for example, be using different words to describe the same issue. This suggests shortcomings in our method of categorizing events, and opportunity for future research.

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

[62]GDELT has been criticized for its low validity and lack of transparency around source data (Wang et al., 2016a), which we observe in our comparison (see the Supplementary Appendix). The Phoenix Near-Real-Time Data produced by the Open Event Data Alliance is projects that works to overcome the shortcomings of GDELT; however, it does not cover the period of our comparison (January to June 2016).

[63]We recognize that the application of CASM, which is trained on social media data, to newspaper data in WiseNews is far from optimal. There many be many more collective action events detailed in WiseNews that we do not capture, but we keep this comparison because the type of collective action events we identify in WiseNews differs from that in CASM-China, even though the application of CASM should bias us toward the identification of similar types of events.

[64]We describe how we collected these data in the Supplemental Appendix.

[65]There are other human-curated protest event datasets, mostly based on newspapers, such as Cai (2010b); Shao (2017). However, none of these datasets are publicly available.

[66]This may change with Social Science One (https://socialscience.one/), but whether researcher can study collective action through Social Science One remains to be seen.

[67]For example, Bruns et al. (2013), Aday et al. (2012), Steinert-Threlkeld et al. (2015), and Steinert-Threlkeld (2017) use Twitter data to study collective action events in the Middle East. González-Bailón et al.

(2011) collect over half a million tweets from Spain. Theocharis et al. (2015) collected Twitter data to analyze protests in Spain, Greece, and the United States.

[68]Won et al. (2017) collect billions of tweets from 14 countries to analyze how images are used by protesters.

REFERENCES

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." Software available from tensorflow.org.

Adams, Nicholas. 2014. "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.

Aday, Sean, Henry Farrell, Marc Lynch, John Sides, and Deen Freelon. 2012. "New media and conflict after the Arab Spring." *Washington: United States Institute of Peace* 80:1–24.

Almeida, Paul and Mark Lichbach. 2003. "To The Internet, From The Internet: Comparative Media Coverage Of Transnational Protests." *Mobilization: An International Quarterly* 8:249–272.

Almeida, Paul D. 2003. "Opportunity Organizations and Threat-Induced Contention: Protest Waves in Authoritarian Settings1." *American Journal of Sociology* 109:345–400.

Azar, Edward E, Stanley H Cohen, Thomas O Jukam, and James M McCormick. 1972. "The problem of source coverage in the use of international events data." *International Studies Quarterly* 16:373–388.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* .

Bamman, David, Brendan O'Connor, and Noah Smith. 2012. "Censorship and deletion practices in Chinese social media." *First Monday* 17.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23:76–91.

Bengio, Yoshua, Ian J. Goodfellow, and Aaron Courville. 2015. "Deep learning." *An MIT Press book in preparation. Draft chapters available at http://www. iro. umontreal. ca/ bengioy/dlbook* .

Bennett, W Lance and Alexandra Segerberg. 2012. "The logic of connective action: Digital media and the personalization of contentious politics." *Information, Communication & Society* 15:739–768.

Blecher, Marc J. 2002. "Hegemony and workers' politics in China." *The China Quarterly* 170:283–303.

Bourgault, Adam. 2015. "Freedom of the Press Under Authoritarian Regimes." *Susquehanna University Political Review* 6:3.

Bruns, Axel, Tim Highfield, and Jean Burgess. 2013. "The Arab Spring and social media audiences: English and Arabic Twitter users and their networks." *American Behavioral Scientist* 57:871–898.

Budak, Ceren and Duncan J Watts. 2015. "Dissecting the Spirit of Gezi: Influence vs. Selection in the Occupy Gezi Movement." *Sociological Science* pp. 370–397.

Cai, Yongshun. 2010a. *Collective Resistance in China: Why Popular Protests Succeed or Fail*. Stanford University Press.

Cai, Yongshun. 2010b. *Collective Resistance in China: Why Popular Protests Succeed or Fail*. Stanford University Press.

Castells, Manuel. 2015. *Networks of outrage and hope: Social movements in the Internet age*. John Wiley & Sons.

Chen, Xi. 2011. *Social protest and contentious authoritarianism in China*. Cambridge University Press.

Chollet, François et al. 2015. "Keras." https://keras.io.

Conneau, Alexis, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. "Very Deep Convolutional Networks for Text Classification." *arXiv:1606.01781 [cs]* arXiv: 1606.01781.

Croicu, Mihai and Nils B Weidmann. 2015. "Improving the selection of news reports for event coding using ensemble classification." *Research & Politics* 2:2053168015615596.

Dahl, George E, Dong Yu, Li Deng, and Alex Acero. 2012. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *IEEE Transactions on audio, speech, and language processing* 20:30–42.

Deibert, Ronald J. 2008. "The geopolitics of internet control: Censorship, sovereignty, and cyberspace." In *Routledge handbook of Internet politics*, pp. 339–352. Routledge.

Deng, Yanhua and Guobin Yang. 2013. "Pollution and protest in China: Environmental mobilization in context." *The China Quarterly* 214:321–336.

Dhingra, Bhuwan, Hanxiao Liu, Ruslan Salakhutdinov, and William W Cohen. 2017. "A comparative study of word embeddings for reading comprehension." *arXiv preprint arXiv:1703.00993* .

Diamant, Neil J. 2010. *Embattled glory: Veterans, military families, and the politics of patriotism in China, 1949–2007*. Rowman & Littlefield Publishers.

Diamond, Larry. 2010. "Liberation technology." *Journal of Democracy* 21:69–83.

Dimitrov, Martin and Zhu Zhang. 2017. "Patterns of Protest Activity in China."

Earl, Jennifer and Katrina Kimport. 2008. "The Targets of Online Protest." *Information, Communication & Society* 11:449–472.

Earl, Jennifer and Katrina Kimport. 2011. *Digitally Enabled Social Change: Activism in the Internet Age*. MIT Press.

Earl, Jennifer, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30:65–80.

Edmond, Chris. 2013. "Information Manipulation, Coordination, and Regime Change." *The Review of Economic Studies* 80:1422–1458.

Egorov, Georgy and Konstantin Sonin. 2011. "Dictators And Their Viziers: Endogenizing The Loyalty–Competence Trade-Off." *Journal of the European Economic Association* 9:903–930.

Ferdinand, Peter. 2000. "The Internet, democracy and democratization." *Democratization* 7:1–17.

Freedom House. 2017. "Press Freedom's Dark Horizon." *Freedom of the Press 2017* .

Gamson, William A. and Andre Modigliani. 1989. "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach." *American Journal of Sociology* 95:1–37.

Goebel, Christian. 2017. "Social Unrest in China A bird's eye perspective."

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. "The Dynamics of Protest Recruitment through an Online Network." *Scientific Reports* 1.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. "The Dynamics of Protest Recruitment through an Online Network." *Scientific Reports* 1.

Guo, Xiaolin. 2001. "Land expropriation and rural conflicts in China." *The China Quarterly* 166:422–439.

Hanna, Alex. 2017. "MPEDS: Automating the Generation of Protest Event Data."

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., X. Zhang, S. Ren, and J. Sun. 2016b. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

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

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal processing magazine* 29:82–97.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9:1735–1780.

Hurst, William. 2004. "Understanding contentious collective action by Chinese laid-off workers: the importance of regional political economy." *Studies in Comparative International Development* 39:94–120.

Hurst, William and Kevin J. O'Brien. 2002. "China's Contentious Pensioners." *China Quarterly* 170:345–360.

Hutter, Swen. 2014a. "Protest Event Analysis and Its Offspring." In *Methodological Practices in Social Movement Research*, edited by Donatella Della Porta. Oxford University Press.

Hutter, Swen. 2014b. "Protest Event Analysis and Its Offspring." In *Methodological Practices in Social Movement Research*, edited by Donatella Della Porta. Oxford University Press.

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

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

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. 2014. "A convolutional neural network for modelling sentences." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

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

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107:326–343.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2014. "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation." *Science* 345:1–10.

Kingma, Diederik P and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* .

Koopmans, Ruud. 2004. "Movements and media: Selection processes and evolutionary dynamics in the public sphere." *Theory and Society* 33:367–391.

Koopmans, Ruud and Dieter Rucht. 2002. "Protest Event Analysis." In *Methods of Social Movement Research*, edited by Bert Klandermans and Suzanne Staggenborg, volume 16, pp. 231–259. University of Minnesota Press.

Koopmans, Ruud and Paul Statham. 1999. "Political claims analysis: integrating protest event and political discourse approaches." *Mobilization: an international quarterly* 4:203–221.

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

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances in neural information processing systems*, pp. 1097–1105.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521:436–444.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1:541–551.

Lee, Ching Kwan. 2007. *Against the Law: Labor Protests in China's Rustbelt and Sunbelt*. University of California Press.

Lee, Sophie J, Howard Liu, and Michael D Ward. 2018. "Lost in Space: Geolocation in Event Data." *Political Science Research and Methods* pp. 1–18.

Liebman, Benjamin L. 2013. "Malpractice mobs: medical dispute resolution in China." *Columbia Law Review* pp. 181–264.

Lorentzen, Peter. 2013. "Regularizing Rioting: Permitting Protest in an Authoritarian Regime." *Quarterly Journal of Political Science* 8:127–158.

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

McAdam, Doug and Yang Su. 2002. "The War at Home: Antiwar Protests and Congressional Voting, 1965 to 1973." *American Sociological Review* 67:696–721.

McAdam, Doug, Sidney Tarrow, and Charles Tilly. 2003. *Dynamics of Contention*. Cambridge University Press.

McCarthy, John D., Clark McPhail, and Jackie Smith. 1996a. "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991." *American Sociological Review* 61:478–499.

McCarthy, John D., Clark McPhail, and Jackie Smith. 1996b. "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991." *American Sociological Review* 61:478–499.

McDonald, Tom. 2016. *Social media in rural China*. UCL Press.

McMillan, John and Pablo Zoido. 2004. "How to Subvert Democracy: Montesinos in Peru." *Journal of Economic Perspectives* 18:69–92.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* .

Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. "Recurrent neural network based language model." In *Eleventh Annual Conference of the International Speech Communication Association*.

Miller, Seumas and Michael J Selgelid. 2007. "Ethical and philosophical consideration of the dual-use dilemma in the biological sciences." *Science and engineering ethics* 13:523–580.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels 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, Taehyun. 2006. "What You Use Matters: Coding Protest Data." *PS: Political Science & Politics* 39:281–287.

Nardulli, Peter F., Scott L. Althaus, and Matthew Hayes. 2015a. "A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* .

Nardulli, Peter F., Scott L. Althaus, and Matthew Hayes. 2015b. "A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* pp. 1–36.

O'Brien, Kevin J. and Lianjiang Li. 2006. *Rightful Resistance in Rural China*. Cambridge University Press.

Oliver, Pamela E. and Gregory M. Maney. 2000. "Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions." *American Journal of Sociology* 106:463–505.

Olzak, Susan. 1989. "Analysis of Events in the Study of Collective Action." *Annual Review of Sociology* 15:119–141.

Ortiz, David, Daniel Myers, Eugene Walls, and Maria-Elena Diaz. 2005. "Where Do We Stand with Newspaper Data?" *Mobilization: An International Quarterly* 10:397–419.

Pan, Jennifer. 2017. "How market dynamics of domestic and foreign social media firms shape strategies of internet censorship." *Problems of Post-Communism* 64:167–188.

Pan, Jennifer and Alexandra Siegel. 2018. "Physical Repression and Online Dissent in the Saudi Twittersphere."

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. 2013. "On the difficulty of training recurrent neural networks." In *International Conference on Machine Learning*, pp. 1310–1318.

Perry, Elizabeth. 2008. "Permanent Revolution? Continuities and Discontinuities in Chinese Protest." In *Popular Protest in China*, edited by Kevin O'Brien, pp. 205–216. Cambridge, MA: Harvard University Press.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep contextualized word representations." In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Qin, Bei, David Stromberg, and Yanhui Wu. 2017. "Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda." *Journal of Economic Perspectives* 31:117–140.

Qin, Bei, Yanhui Wu, and David Strömberg. 2012. "The determinants of media bias in China." Working Paper.

Rainie, Lee, Aaron Smith, Kay Lehman Schlozman, Henry Brady, and Sidney Verba. 2012. "Social media and political engagement." 19.

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David 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, pp. 1799–1808, New York, NY, USA. ACM.

Rattani, A. and R. Derakhshani. 2017. "On fine-tuning convolutional neural networks for smartphone based ocular recognition." In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 762–767.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2015. "Faster r-cnn: Towards real-time object detection with region proposal networks." In *Advances in neural information processing systems*, pp. 91–99.

Roberts, Margaret E. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press.

Romero, Daniel M, Brendan Meeder, and Jon Kleinberg. 2011. "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter." In *Proceedings of the 20th international conference on World wide web*, pp. 695–704. ACM.

Rucht, Dieter, Ruud Koopmans, and Friedhelm Neidhardt. 1999. *Acts of dissent: new developments in the study of protest*. Rowman & Littlefield.

Sainath, T. N., O. Vinyals, A. Senior, and H. Sak. 2015. "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584.

Sak, Haşim, Andrew Senior, and Françoise Beaufays. 2014. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." In *Fifteenth annual conference of the international speech communication association*.

Santos, Cicero D and Bianca Zadrozny. 2014. "Learning character-level representations for part-of-speech tagging." In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1818–1826.

Saraf, Parang and Naren 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, pp. 599–608, New York, NY, USA. ACM.

Schuster, M. and K. K. Paliwal. 1997. "Bidirectional recurrent neural networks." *IEEE Transactions on Signal Processing* 45:2673–2681.

Selgelid, Michael J. 2013. "Dual-Use Research." In *International Encyclopedia of Ethics*. American Cancer Society.

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

Shin, Hoo-Chang, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2497–2506.

Simonyan, Karen and Andrew Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* .

Simonyan, Karen and Andrew Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In *Proceedings of the Third International Conference on Learning Representations*.

Smith, Aaron. 2013. "Civic Engagement in the Digital Age." Pew Research Center.

Steinert-Threlkeld, Zachary C. 2018. *Twitter as Data*. Cambridge University Press.

Steinert-Threlkeld, Zachary C. 2017. "Spontaneous collective action: peripheral mobilization during the Arab Spring." *American Political Science Review* 111:379–403.

Steinert-Threlkeld, Zachary C, Delia Mocanu, Alessandro Vespignani, and James Fowler. 2015. "Online social networks and offline protest." *EPJ Data Science* 4:19.

Stern, Rachel E and Jonathan Hassid. 2012. "Amplifying silence: uncertainty and control parables in contemporary China." *Comparative Political Studies* 45:1230–1254.

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

Su, Yang and Xin He. 2010. "Street as courtroom: state accommodation of labor protest in South China." *Law & Society Review* 44:157–184.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. "Sequence to sequence learning with neural networks." In *Advances in neural information processing systems*, pp. 3104–3112.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. "Going Deeper With Convolutions." pp. 1–9.

Tan, Chade-Meng, Yuan-Fang Wang, and Chan-Do Lee. 2002. "The use of bigrams to enhance text categorization." *Information processing & management* 38:529–546.

Tarrow, Sidney. 2005. *The New Transnational Activism*. Cambridge University Press.

Theocharis, Yannis, Will Lowe, Jan W Van Deth, and Gema García-Albacete. 2015. "Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements." *Information, Communication & Society* 18:202–220.

Tilly, Charles. 2008. *Contentious performances*. Cambridge University Press.

Tong, Yanqi and Shaohua Lei. 2010. "Large-scale mass incidents and government responses in China." *International Journal of China Studies* 1:487–508.

Torres, Michelle. 2018. "Give me the full picture: Using computer vision to understand visual frames and political communication." *Working Paper* .

Trentham, Barry, Sandra Sokoloff, Amie Tsang, and Sheila Neysmith. 2015. "Social media and senior citizen advocacy: an inclusive tool to resist ageism?" *Politics, Groups, and Identities* 3:558–571.

Wang, Wei, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016a. "Growing pains for global monitoring of societal events." *Science* 353:1502–1503.

Wang, Xingyou, Weijie Jiang, and Zhiyong Luo. 2016b. "Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2428–2437, Osaka, Japan. The COLING 2016 Organizing Committee.

Weiss, Jessica Chen. 2014. *Powerful patriots: nationalist protest in China's foreign relations*. Oxford University Press.

Wilson, Ashia C, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. 2017. "The marginal value of adaptive gradient methods in machine learning." In *Advances in Neural Information Processing Systems*, pp. 4148–4158.

Won, Donghyeon, Zachary C Steinert-Threlkeld, and Jungseock Joo. 2017. "Protest Activity Detection and Perceived Violence Estimation from Social Media Images." In *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 786–794. ACM.

Xiao, Yijun and Kyunghyun Cho. 2016. "Efficient Character-level Document Classification by Combining Convolution and Recurrent Layers." *arXiv:1602.00367 [cs]* arXiv: 1602.00367.

Yang, Guobin. 2003. "The Internet and the rise of a transnational Chinese cultural sphere." *Media, Culture & Society* 25:469–490.

Zhou, Chunting, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. "A C-LSTM Neural Network for Text Classification." *CoRR* arXiv: 1511.08630.

Zhu, Tao, David Phipps, Adam Pridgen, Jedidiah R Crandall, and Dan S Wallach. 2013. "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions." In *USENIX Security Symposium*, pp. 227–240.

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

Abstract

TBD

1 Collective Action Coding Rules

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

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

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

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

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

2 Selecting Keywords K

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

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

repression; legal rights; appeal; law of the land; defraud ; truth;children; forcefully take land; redress an injustice; lawyers; petitioners; be responsible for the people ; maintain stability; sit-in; forced land taking

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

Note that some Chinese terms have the same English translations.

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

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

We also evaluate the robustness of CASM's output to size of K. To do that, we create a subset of T_K that includes the top n keywords in K, and see how the output of CASM-China is impacted by the increase of k. Figure 4 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 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

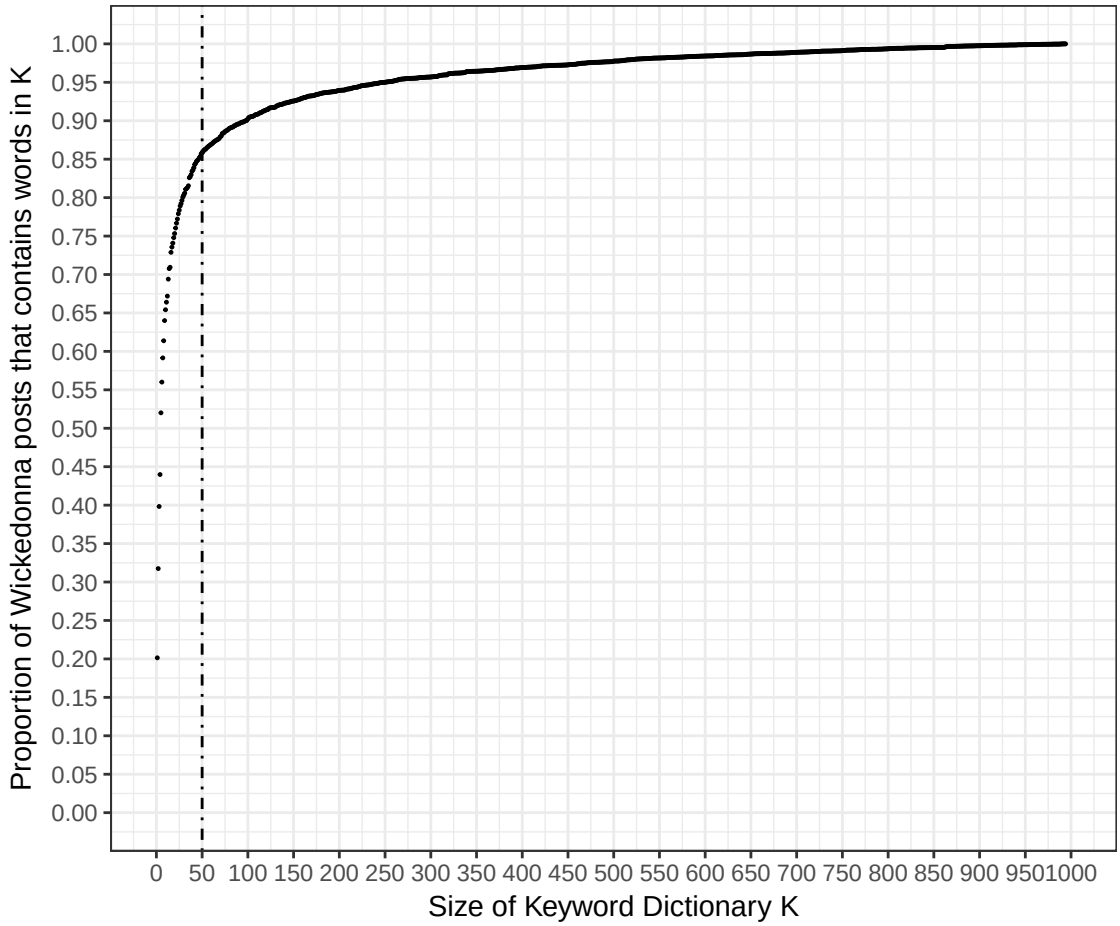


Figure 1: Coverage of protests in the Wickedonna Dataset by size of keyword dictionary; beyond 50 keywords, marginal coverage declines.

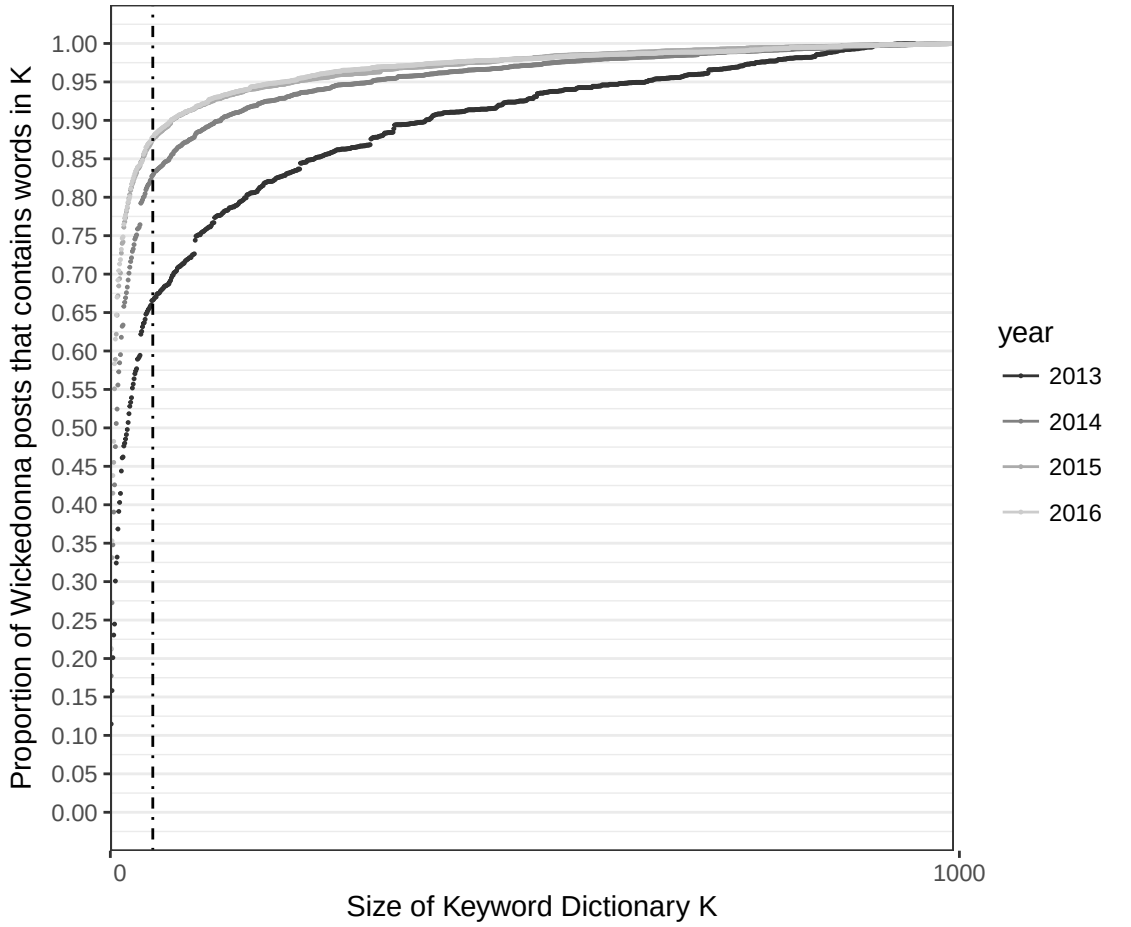
current size of 50 is unlikely to substantially impact the identification of collective action events by CASM.

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

3 Model Comparison

[*Maybe remove this altogether, or add more details. Update response to Reviewer 2 –JP*][*I think we should keep it. People will love to see the performance comparisons. –HZ*]. Figure 6 shows how our CNN-RNN deep learning model (solid line) outperforms conventional supervised machine learning algorithms. We compare the CNN-RNN model from the first-stage with SVM and Naive Bayes. The exact same pre-processed training data is used The comparison between the second-stage classifier, SVM and Naive Bayes shows a similar trend. [*Need more details on SVM and Naive Bayes - specifically, what's done in pre-processing*

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



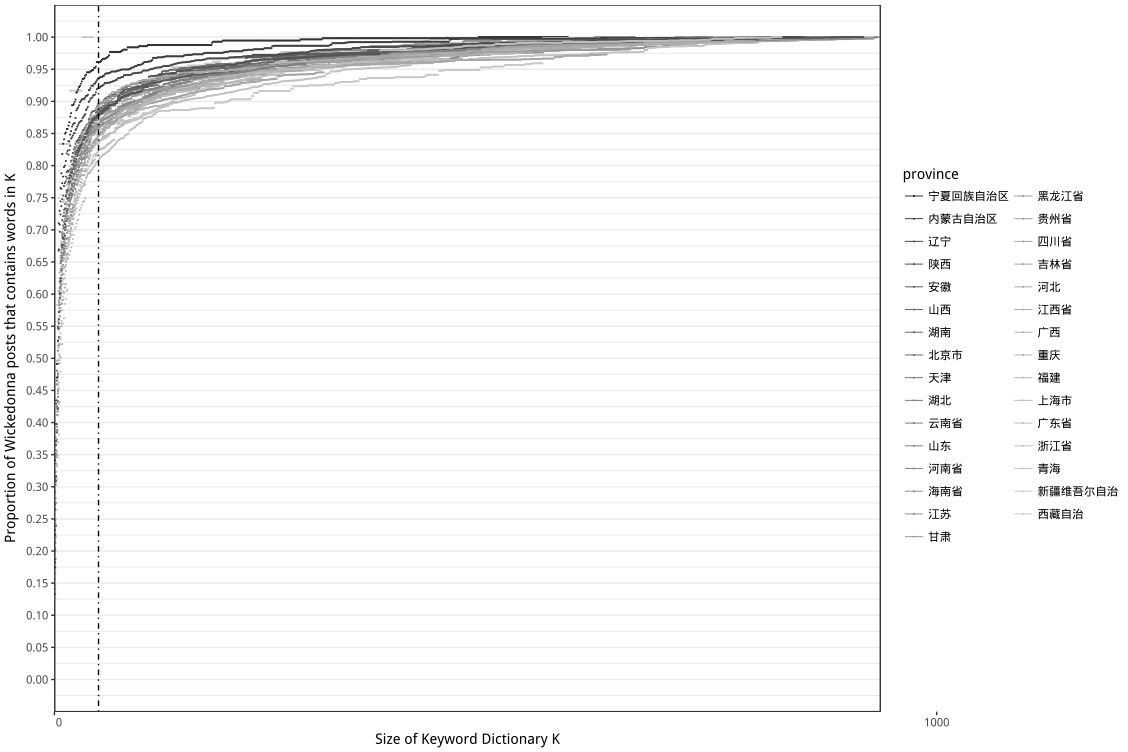
–JP][*The preprocessing used the same stopword, and same segmentation algorithm. We use n-gram with n ranging from 1 - 5. We remove words that appear less than 5 times. We also used tf-idf transformation of term-document matrix which weigh down frequent words while scale-up rare ones. –HZ*]

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

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

- Conventional: “parade”, “strike”, “assembly”, “protest”, “voluntarily”, “upper level petition”, “name list”, “defend rights”, “petition”, “asking for back wages”, “ar-

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



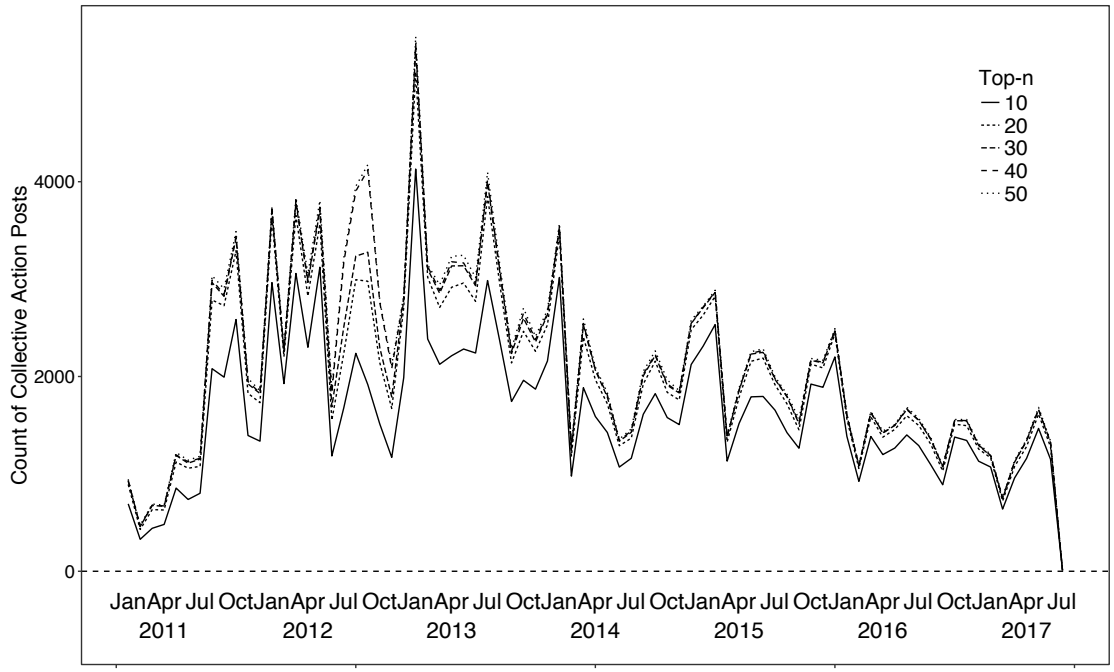


Figure 4: The number of collective-action events identified in CASM-China by the size of the dictionary.

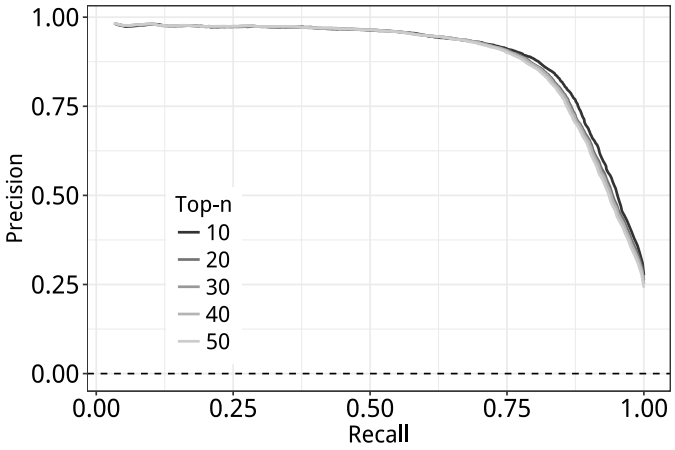


Figure 5: Precision-recall curve by size of keyword dictionary

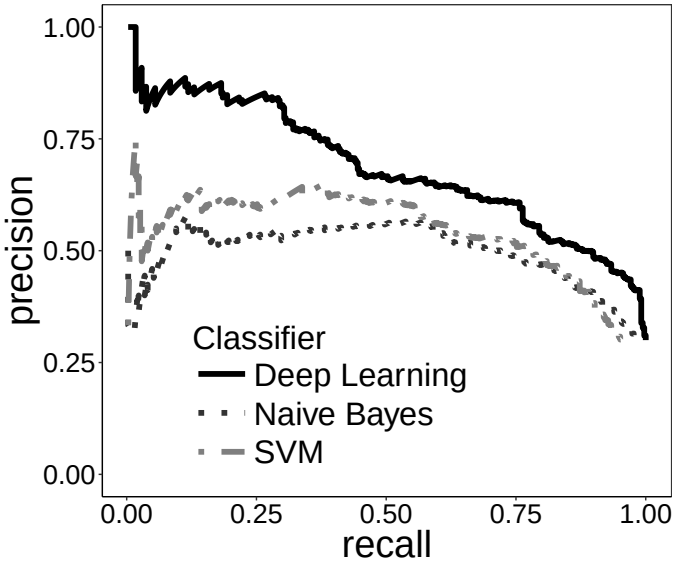


Figure 6: Precision-recall curve: deep learning vs. conventional algorithms. The comparison is based on texts only.[1]

rears", "arrears", "pleading for help", "hard-earned money", "protect", "interest" ("游行", "罢工", "集会", "抗议", "情愿","上访", "签名", "维权", "信访", "讨薪", "欠薪", "拖欠", "求助", "血汗钱", "维护", "利益')

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

- Violent: "attack", "violence", "armed police", "police", "beat down", "police", "forced demolition", "forced", "police vehicle", "forced taking" "攻击", "暴力", "特警", "警察", "打倒","警务","强拆", "强行", "警车","强征"

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

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

2. Ethnic / religious: “Christianity”,“Catholism”,“Protestantism, ”church”, “Xinjiang”, “Uighur”, “Uighur”, Islam”, “Muslim”, “Hui Muslims”, “Tibet”, “Tibetan”, “Tibetans’, “self-immolation”, “Islam” “基督教”, “天主教”, “新教’, ”教堂”, “新疆”, “维族”, “维吾尔”, “绿教”, “穆斯林”, “回族”, “西藏”, “藏族”, “藏民”, “自焚”, “伊斯兰”

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

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

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

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

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

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

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

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

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

5 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 **??**. 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.[2]. Similar to GDELT, ICEWS also monitors global news agencies, but places more emphasis on the accuracy of identifying events rather than documenting as many events as possible (Ward et al., 2013). We first extract events between January to June 2016 that 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.[3]

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

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

Wickedonna: We introduced the Wickedonna Dataset in Section **??**. 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 our 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.[6] 14% of the 500 events (7.3% =52%* 0.4 of the total population) are no longer

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

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

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

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

[6]Note that 16.6% is based on the sample of 500 events. Earlier we found that 14% of all Wickedonna events lacked one of the 50 keywords.

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[7] betwen June 2013 to June 2016, which accounts for 53.3% of their entire dataset. For the period between January to June 2016, 81% of events in the China Labor Bulletin are from the Wickedonna dataset. We coded a random sample of China Labor Bulletin strikes (200 events) and find that 75% of their events are in CASM-China.

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

References

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances in neural information processing systems*, pp. 1097–1105.

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

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