Combining Heterogeneous Data Sources for Civil Unrest Forecasting

Gizem Korkmaz*, Jose Cadena*†, Chris J. Kuhlman*, Achla Marathe*, Anil Vullikanti*†, Naren Ramakrishnan†
*Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24060, USA
†Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA
Email: {gkorkmaz,jcadena,ckuhlman,amarathe,akumar}@vbi.vt.edu, naren@vt.edu

***Abstract*—Detecting and forecasting civil unrest events (protests, strikes, etc.) is of key interest to social scientists and policy makers because these events can lead to significant societal and cultural changes. We analyze protest dynamics in six countries of Latin America on a daily level, from November 2012 through August 2014, using multiple data sources that capture social, political and economic contexts within which civil unrest occurs. We use logistic regression models with Lasso to select a sparse feature set from our diverse datasets, in order to predict the probability of occurrence of civil unrest events in these countries. The models contain predictors extracted from social media sites (Twitter and blogs) and news sources, in addition to volume of requests to Tor, a widely-used anonymity network. Two political event databases and country-specific exchange rates are also used. Our forecasting models are evaluated using a Gold Standard Report (GSR), which is compiled by an independent group of social scientists and experts on Latin America. The experimental results, measured by F1-scores, are in the range 0.68 to 0.95, and demonstrate the efficacy of using a multi-source approach for predicting civil unrest. Case studies illustrate the insights into unrest events that are obtained with our methods.**

I. Introduction

A. Background and Motivation

Civil unrest events unfold through complex mechanisms that are not fully understood. Factors in the emergence of civil unrest include social interactions and injustices, changes in domestic and international policies cultural awareness, and economic factors, such as poverty, unemployment levels, and food prices [1], [2], [3]. What is clear is that protests and social upheaval, even if small and non-violent at first, have the potential to evolve into nationwide events [2]. Predicting the occurrence of civil unrest events is in the interest of policy makers, since local unrest can lead to regional instability [4].

In recent years, open source data, such as social media content, have been used with varying degrees of success to forecast civil unrest. One limitation of many such studies is that the methods are optimized for harnessing data from a single source (see Section II). This approach has several disadvantages. If the data source in question becomes unavailable, it is unclear whether or how quickly the models can be adapted to new alternative sources. Furthermore, civil unrest events are complex processes that cannot be fully characterized by looking at one feed in isolation (we provide one example from among many herein in which Twitter—a popular data source—misses a protest event). We posit that combining different types of indicators of civil unrest, such as social media, political and opinion blogs, news sources, and measures of economic performance creates a more informed and robust signal that can forecast civil unrest events. Recent events such as mass protests in the Middle East (Turkey, Egypt, Tunisia) and South America (Brazil, Venezuela) provide anecdotal evidence of the value that can be created by aggregating different sources. During these events, even when authorities tried to censor one source of information, such as Twitter, demonstrators found ways to voice their dissent and concerns through alternate media and networking outlets, e.g., using anonymity networks like Tor.[1]

B. Contributions

We present a model for predicting civil unrest through the combination of heterogeneous online data sources. To the best of our knowledge, this is the first model of social unrest forecasting that combines several relevant data sources, and critically evaluates the approach.[2] Our main contributions follow.

1. Forecasting of civil unrest events using social, economic and political indicators. We develop statistical models (Lasso and hybrid models) to forecast civil unrest events in six Latin American countries using multiple data sources. Specifically, features from social media (Twitter, blogs), news sources, two political event databases, Tor statistics, and exchange rates are combined. Our predictions are compared to a longitudinal data set of civil unrest events in Latin America, as identified by an expert panel, over the course of 2 years. We show that combining heterogeneous data sources is effective in predicting next-day civil unrest, as measured by the F1-score (which is a measure of the quality of the forecasts that balances precision and recall). Our model produces F1-scores in the range 0.68 to 0.95, which compare well with the overall system [5] that fuses multiple prediction models including our model, with F1-scores in the range 0.62 to 0.83, and lead times up to 8 days. Our performance results are provided in Section V-B. Lasso selects different information sources, across countries, for their predictive value. A baseline model that uses only past ground truth data in predicting events does fairly well because there are many events in these countries. However, as opposed to the Lasso model, it does not give any information about the nature of particular social unrest events.

[1] June 21, 2013, "Protesters, criminals get around government censors using secret web network," http://bit.ly/1Sghvo7.

[2] This model was mentioned, along with several others in [5] as part of an automated, real-time forecasting software system. This paper is the first to describe our model and results in detail.

2. Interpretable feature selection supported by case studies. Lasso-based models provide analysts with insights about the underlying social dynamics in different countries by identifying predictive features that are tied to unrest. This is illustrated by the case studies in Section V-C. For Brazil, protests about financial concerns translate into variations in the exchange rate, which in turn result in forecasts of higher probabilities of civil unrest. Moreover, the frequent use of the word "racismo" in Twitter messages turns out to be a harbinger of a racism-motivated protest. Similarly, "represión" was selected by Lasso prior to protests in Venezuela. Supporting these data was the increased Tor activity, through which people can act online anonymously (due to strict controls of the Venezuelan government over demonstrators). The case studies show that different signals can be obtained from different data sources and that features picked by Lasso provide additional information about various unrest events.

3. Value of data sources. Our ground truth data—the Gold Standard Report (GSR)—are produced by an independent panel of experts on Latin America and describe civil unrest events in detail (Section III-A). Our study provides the following three observations about the data sources. The first one relates to Twitter. There are many studies that use Twitter to identify social unrest [6], [5]. Here, we provide an example of a 2014 protest event in Venezuela in which blog activity identifies an outburst of protest events but Twitter activity does not (Figure 3). Second, Tor and exchange rate were each found to be significant predictors of particular events in different countries. These findings illustrate the benefits of using multiple data sources (Section V-C). The third finding is surprising in that political event databases do not provide particularly strong signals of protests and civil unrest. This is more surprising because unrest events, according to the GSR, predominantly stem from government policies (Figure 5).

II. Related Work

There is a significant amount of prior work on detecting and predicting real-world events using social media data. Twitter, in particular, has received a lot of attention. Data from this social media site have been used to predict events as diverse as movie box-office revenue [7], political elections [8], the stock market [9], flu trends [10], [11], [12], and even earthquakes [13]. A summary of the different predictive tasks studied and the proposed methods can be found in [14]. More recently, Twitter has also been used to forecast civil unrest [6], [5]; however, most of the previous work involving Twitter has focused on how people interact on the social media site in times of protest [15], [16], [17] and far less on event forecasting.

The other data sources that we consider in this paper have been used for event prediction to a lesser extent. In [18], Kallus uses data from news and blogs articles and tweets to train a random forest classifier to predict big-scale protest events in 18 countries. The authors in [19] use news data to address the related problem of predicting conflict between countries.

A non-trivial task in the process of forecasting an event using news and blog sites is mining articles and determining which of these are relevant for the prediction task. An alternative is to use political event datasets, such as ICEWS [20] and GDELT [21]. These datasets are daily compilations of events extracted from news reports around the world. The events are automatically coded for type of event (conflict or cooperation), entities involved (countries, state heads, military, etc.), and severity of the event. ICEWS and GDELT have been used to forecast large-scale political conflict events, such as rebellion, insurgency, domestic crises, and international crises [22], [23].

To the best of our knowledge, the remaining two data sources that we consider—Tor usage metrics and currency exchange rates—have not been used as predictors of civil unrest or related topics. Furthermore, we note that our methodology differs from the previous works described above in that we combine data from many different sources, thus capturing different aspects of a civil unrest event. With the exception of [12], which uses variants of matrix factorization to combine data from 7 sources including Twitter, Google Trends, weather etc. to forecast flu activity, all the methods above consider only one or two datasets for prediction. The focus of [5] was the fusion of predictions from multiple *models* for realtime forecasting. One of those models [24] uses multiple sources, individually, to identify protests that are called out in different media.

The Lasso regression [25] has garnered interest in diverse fields in recent years. In social media analytics, specifically, Lasso has been used to train models for predicting elections [26] and detecting flu with Twitter data [11]. One application of Lasso in social media with a similar goal to ours (i.e. combining different data sources) can be found in [27], where the authors integrate data from the social media sites BlogCatalog and Flickr for community detection.

III. Data Sources

A. Gold Standard Report (GSR)

The GSR dataset is a compilation of occurrences of civil unrest in Latin American countries from November 1, 2012 to August 31, 2014, and serves as the ground truth for our evaluation. The events are manually extracted from well-reputed newspapers for each country by an independent group of social scientists and experts in Latin American politics. It has information about the exact location and the date of the event as well as the date when it was reported in a news source. Not only are the events coded based on the purpose of the unrest (e.g., 014-Wages and Employment, 013-Energy and Resources, etc.), but also the events are associated with a "population type" based on the demographic of the people involved in the insurgency (e.g., Labor, Education, etc.). Our focus is forecasting the first day of nationwide, relatively larger protests that span multiple locations, i.e., multiple cities and states within the country. In the GSR, these events fall under "General Population" (i.e., people from diverse demographics).

Table I shows a partial entry from the GSR about a nationwide protest that took place in July 15, 2014 in Mexico. As shown in the table, every entry in the GSR includes the date and location of the event, the event type and population type. Additional information, not shown, includes a URL to the news article where the event was first reported. Finally, Figure 1 and the accompanying table show the frequency of events in Latin America. In this paper, we focus on the six countries with

TABLE I: An example of an event in the GSR

Country	State	City	Population	Event Type	Date	Source
Mexico	-	-	General Population	013 - Energy and Resources	2014-07-15	Milenio

Country	Events
Mexico	4,454
Venezuela	3,072
Brazil	3,051
Paraguay	1,800
Argentina	1,227
Colombia	1,069
Chile	638
El Salvador	608
Uruguay	570
Ecuador	410

Fig. 1: Numbers of GSR events across countries in Latin America (November 1, 2012 - August 31, 2014). Mexico, Venezuela and Brazil have the highest number of events in Latin America.

the highest volume of civil unrest events (Argentina, Brazil, Mexico, Colombia, Paraguay and Venezuela).

B. Data Sources for Forecasting Models

Social media, blogs, and news. Social media activity is captured by mining Twitter, news websites, and popular blogs of each country. We use a Twitter dataset of around 500 million tweets, encompassing the period of November 2012 to August 2014. The data, obtained from Datasift, constitutes a 10% sample of the tweets emitted in Latin America for the period. Table II reports a summary of the number of tweets for each country for the duration of the study. Our news datasets consist of articles collected from news sources obtained from the Latin American Network Information Center (LANIC)[3] and websites found in online newspaper repositories. Articles were collected by subscribing to the RSS feeds of these sites. In total, articles from 6,236 news websites are collected. Similar to the news dataset, we subscribed via RSS to popular blog sites in Latin America. In total, we gathered blog posts from 3,262 websites.

Civil unrest activity in social media is summarized by filtering and aggregating our datasets using a dictionary of 962 protest-related keywords that was compiled by political

[3]http://lanic.utexas.edu/

TABLE II: Number of collected tweets for each country (November 1, 2012 - August 31, 2014)

Country	Number of Tweets
Brazil	192,412,090
Argentina	130,348,578
Venezuela	115,850,700
Mexico	114,037,524
Colombia	71,258,239
Paraguay	13,598,093

Fig. 2: Word cloud of the civil unrest dictionary used for filtering Twitter, blogs, and news. It includes *protest-related words* ("gobierno" (government), "estudiantes" (students)), *key phrases* ("salir a la calle" (take the streets)), country-specific *key players* ("Henrique Capriles").

scientists and domain experts on Latin America.[4] The dictionary, which is illustrated as a word cloud in Figure 2, includes *keywords* that could be (i) *protest words* such as protesta, revolución, marcha, (ii) *key phrases* such as "marcha por la paz" (walk for peace), or (iii) country-specific *key players* (political parties, heads of state, unions) such as "Henrique Capriles" (leader of the opposition party in Venezuela). Translations of the keywords in Spanish, Portuguese and English are also used for filtering.

From each of our social media datasets, we sub-select the tweets, news, and blog articles that contain at least three different keywords from this dictionary. This subset is considered to pertain to civil unrest. The keyword volumes for one day are a proxy for the amount of protest-related chatter for that day. Specifically, for each one of these datasets, daily counts of keywords associated with civil unrest are computed and used as features for civil unrest prediction.

As an example of the relationship between social media activity and real-world protests, the time series of keyword counts on Twitter and blogs are compared to the time series of GSR events in Figure 3. The time series of GSR is superimposed to Twitter (left) and blogs (right); each series is normalized by its respective maximum count. The daily Twitter volumes capture the general trend of GSR events. Particularly, both series peak during February and March 2014, a month where the country was in a constant state of nationwide protests. The Twitter time series "miss" the peak of GSR events around May 2014. However, the surge in events is picked up by the activity in blogs. The above is a descriptive example, out of many, which shows that using different sources of social media data in combination can give better signals about civil unrest events than using the sources in isolation, capturing different events.

Political event databases. Political event databases are formed by identifying interactions between political *actors*—this can be as general as a country or an ethnic group, or as specific as a particular person—from news sources throughout the world. These datasets are of immense value in the social sciences and they have been used for the prediction of large-scale demonstrations and political crises (see Section II). Two such datasets are considered in our study: Integrated Crisis Early

[4]The dictionary is compiled by a different group of experts from those that generated the GSR.

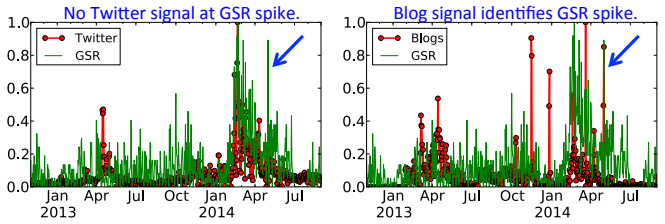

Fig. 3: Normalized time series of keyword counts in Twitter (left) messages and blog posts (right) are compared to GSR events (in green) in Venezuela. The peak of events around May 2014 is picked up by the activity in blogs, but is "missed" by Twitter.

TABLE III: CAMEO Code Summary

Cooperation	Conflict
Verbal	
01 - Make public statement	11 - Disapprove
02 - Appeal	12 - Reject
03 - Express intent to cooperate	13 - Threaten
04 - Consult	14 - Protest
05 - Engage in diplomatic cooperation	15 - Exhibit Force Posture
Material	
06 - Engage in material cooperation	16 - Reduce relations
07 - Provide aid	17 - Coerce
08 - Yield	18 - Assault
09 - Investigate	19 - Fight
10 - Demand	20 - Use unconventional mass violence

Warning System (ICEWS) [20] and Global Data on Events, Location and Tone (GDELT) [21]. Both of these databases classify events through a CAMEO coding convention [28], so each event is assigned one of 20 categories that indicates whether the interaction is collaboration or conflict and whether the interaction is material or verbal in nature. Table III shows how the 20 CAMEO categories are distributed among the cooperation / conflict and material / verbal dimensions.

For each of the 20 CAMEO event types, the following features are used in our model: (i) daily counts of events, (ii) average intensity of the events in ICEWS, (iii) average tone of the daily events in GDELT (aimed to measure the general sentiment of the entities involved in the event), and (iv) Goldstein scale score of daily events in GDELT (a collaboration score assigned to each event; the higher the score between the two actors, the greater their collaboration).

Tor. The Onion Router[5] (Tor) is a free anonymization tool that protects an individual's identity and location on the Internet. The Tor daily usage statistics could be an indicator of preparations for uprising, especially in suppressed societies. The daily volume of requests to Tor is another feature of our model.

Currency. To capture the economic stability of countries, country-specific currency exchange rates against the dollar are collected from Yahoo! Finance. The exchange rate is used as a proxy for the economic performance of the country.

IV. Proposed Methods

Our goal is to develop a model for forecasting civil unrest events by combining all the data sources described above. We pose this forecasting task as a classification problem: Let $\mathbf{X}_t \in \mathbb{R}^p$ be a *feature vector* that summarizes the data corresponding to a day t, and let $y_{t+1} \in \{0, 1\}$ be an indicator variable that has value 1 if there is a civil unrest event on day $t + 1$. The goal is to find a function $f : \mathbb{R}^p \rightarrow \{0, 1\}$, such that

$$f(\mathbf{X}_t) = y_{t+1}.$$

At the same time, we are interested in finding which features are good predictors of protest in the countries we study. For the data under consideration, we use the GSR on $n = 669$ days and a total of $p = 2,988$ features from the data sources described above, all of which have a granularity of one day (see Table IV). However, it is unlikely that all these features are good signals of civil unrest. Given the socioeconomic, political and cultural differences across the six countries under consideration, it is also not the case that there is a *one-size-fits-all* set of features that produces good forecasts for all the countries. Therefore, feature selection methods are used to reduce the dimensionality of our data.

[5]Source: https://www.torproject.org/

TABLE IV: Number of Features Per Data Source

Data Source	Num. Features	Data Source	Num. Features
Twitter	962	ICEWS	
News	962	Events	20
Blogs	962	Avg. Int.	20
Tor	1	GDELT	
Currency	1	Events	20
GSR	1	Avg. Tone	20
		Goldstein	20

A. Logistic regression with Lasso regularization

Since the features of our model outnumber observations ($p >> n$), we employ Least Absolute Shrinkage and Selection Operator (Lasso) regression for feature selection [25]. Lasso is a penalized likelihood method for model estimation that performs simultaneous variable selection and coefficient estimation to produce a parsimonious list of predictors. Lasso is a constrained version of ordinary least squares (OLS) regression (Lasso minimizes the sum of squared errors, but with an added constraint on the sum of the absolute values of the coefficients), and it computes a sparse regression estimate vector, β^*, by solving the following optimization problem:

$$\beta^* = \text{argmin}_\beta \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is a *feature matrix* (i.e., each row corresponds to a feature vector), $\mathbf{Y} \in \{0, 1\}^n$ is binary vector of response variables, and $\beta \in \mathbb{R}^{p+1}$ is a vector of coefficients to be estimated (p features and one intercept term). $\lambda \geq 0$ is the regularization or shrinkage parameter, since it controls the penalty on the L1-norm of β's; high values of the parameter encourage sparser models by reducing many of the coefficients to zero and hence the Lasso method can perform the variable selection procedure. The regularization parameter is determined by cross-validation.

For our prediction task, we run a logistic regression with Lasso penalty, since we use a binary dependent variable. Formally, we estimate

$$logit(P_{t+1}) = log\left(\frac{Pr(y_{t+1} = 1)}{1 - Pr(y_{t+1} = 1)}\right) = \mathbf{X}_t\boldsymbol{\beta},$$

where y_{t+1} denotes the GSR characterization of civil unrest events, which is a binary variable: 1, if there is an event on day $t+1$ in the GSR; 0 otherwise. Similarly, $\mathbf{X}_t$ is the feature vector for day t; each element of $\mathbf{X}_t$ represents a summary of the various data sources discussed in Section III.

The Lasso method was chosen because the number of features that can be used to train the model is much larger than the number of observations, so we have to find a reduced set of features first. There are many dimensionality-reduction techniques based on singular value decomposition [29]. However, the reduced feature space returned by these methods does not have an intuitive interpretation. By using Lasso, we simultaneously train a predictive model and obtain a reduced set of features that are not only related to civil unrest, but are also interpretable.

There is an extension to Lasso called *group-Lasso* [30]. This model shrinks entire groups of coefficients to zero. We evaluated two models based on group-Lasso: (i) variables are grouped based on the data sources; thus, we use 7 groups: Twitter, news, blogs, Tor, currency, ICEWS, and GDELT; (ii) text-based data sources (Twitter, news and blogs) are grouped based on the specific keyword. For example, the counts of the keyword "protest" in Twitter, news and blogs are defined as one group. The other features (that are not text-based) are not grouped and are used independently. Both of these methods performed poorly compared to Lasso; thus, we omit the details here.

B. Baseline Model

Large-scale civil unrest events are processes that grow and gain momentum over several consecutive days. Therefore, the occurrence of an event on a given day can be used as a predictor of the occurrence of an event in the near future. In order to capture this serial correlation structure, we use a baseline model that uses no external input in the regression model. Rather, it uses lagged values of the GSR (whether an event occurred in the previous day) as the sole predictor of protest in a given day. The binary prediction $\hat{y}_{t+1}$ for an event is given by

$$\hat{y}_{t+1} = \begin{cases} 0 & \text{if } y_t = 0 \\ 1 & \text{if } y_t = 1 \end{cases}$$

This model sets a benchmark in order to measure the added predictability provided by our datasets.

C. Hybrid Model

We train a *hybrid* model that combines the lagged value of the GSR with the information obtained from our various datasets. Formally, the model estimates:

$$\hat{P}_{t+1} = log\left(\frac{Pr(y_{t+1}=1)}{1-Pr(y_{t+1}=1)}\right) = y_t + \mathbf{X}_t\boldsymbol{\beta}.$$

The hybrid model uses a Lasso logistic regression for feature selection, but it also imposes a mandatory AR(1) structure that is independent of the Lasso selection.

V. Experiments

Our experiments are designed (i) to evaluate the robustness of the predictions of our models for different training periods, and over different periods of time across the whole dataset, (ii) to identify the relevant features in capturing events of different natures, and (iii) to compare the predictive performance of our models against the baseline in terms of F1-score (defined below) and their ability to provide insights about the social dynamics in the countries using case studies.

We begin by describing our evaluation methodology and performance metrics in the next section.

A. Experimental Methodology

Preprocessing. Before training the models, every variable (i.e., every column in the feature matrix) is standardized to have mean zero and unit variance.

Metrics. The performance of the models is evaluated using standard classification metrics: (i) Recall (True Positive Rate): the percentage of GSR events that the model predicts; (ii) Precision: the percentage of event predictions of the model that are actually matched with a GSR event; and (iii) F1-score: the weighted harmonic average of precision and recall.

Training and testing. In order to assess the robustness of the models, different training periods of 300, 350, and 400 days are used. The ten-fold cross validation approach is used on the training sets to tune the regularization parameter λ. After training each model, predictions are made for the testing set, i.e., the next T days following the training set, where $T \in \{10, 20, 30\}$.

Given the Lasso estimate for the probability of an event, $\hat{P}_{t+1}$ on day $t+1 \in T$, the following prediction is made for each day over T:

$$\hat{y}_{t+1} = \begin{cases} 0 & \text{if } \hat{P}_{t+1} \leq \tau^* \\ 1 & \text{if } \hat{P}_{t+1} > \tau^* \end{cases}$$

where the threshold τ^* is determined by maximizing the F1-score over T by comparing the predictions, $\hat{y}_{t+1} \in \{0, 1\}$, to the actual events in the GSR. The precision, recall, and F1-scores are then computed for measuring the model's performance for the testing set.

The training set is shifted by 5 days after the evaluation of the model. The training and the evaluation processes are repeated for the new training and testing sets. Hence, we obtain performance metrics for multiple testing sets across the whole dataset.

B. Performance Results

Table V summarizes the predictive performance of the regularized logistic regression for all countries. The table reports—from left to right—the size of the testing set, T, the number of new GSR events (to be predicted) averaged across the testing sets, and the performance metrics (i.e., precision, recall, and the F1-score); due to space constraints, only the average values over the multiple testing sets are reported in the table. The average recall is high in all of the countries (ranging from 0.94 to 1.0) and the precision ranges from 0.55

TABLE V: Performance metrics of Lasso for 300 training days.

Country	Test Days	Average # Events	Precision	Recall	F1-score
Argentina	$T = 10$	4.84	0.55	0.97	0.70
	$T = 20$	9.18	0.55	0.97	0.70
	$T = 30$	12.95	0.55	0.94	0.69
Brazil	$T = 10$	7.83	0.80	1.00	0.89
	$T = 20$	16.00	0.89	1.00	0.94
	$T = 30$	23.68	0.91	1.00	0.95
Colombia	$T = 10$	4.89	0.51	1.00	0.68
	$T = 20$	9.38	0.53	0.99	0.69
	$T = 30$	14.17	0.55	0.99	0.70
Mexico	$T = 10$	7.55	0.78	1.00	0.88
	$T = 20$	15.11	0.77	0.97	0.86
	$T = 30$	22.08	0.78	0.98	0.87
Paraguay	$T = 10$	5.64	0.58	0.96	0.72
	$T = 20$	11.16	0.56	0.98	0.71
	$T = 30$	17.17	0.57	0.97	0.72
Venezuela	$T = 10$	6.46	0.68	0.99	0.80
	$T = 20$	15.11	0.69	0.97	0.81
	$T = 30$	21.00	0.71	0.98	0.83

*The results are similar for the hybrid model.

to 0.91. Precision is lower for Argentina and Colombia, which are the countries with fewer numbers of events.

In Table V, we observe that the average F1-scores are in the range 0.68 to 0.95 and their distributions are illustrated in Figure 4. Each boxplot represents the distribution of the scores obtained for each (model, testing period) pair (e.g., Baseline $T = 20$ indicates that the baseline model is used and the testing period is 20 days throughout the experiment). The variation in the performance (due to the shifting of the training and testing periods) is captured by the height of the box, and the horizontal lines within the boxes indicate the means of the scores. Due to space limitations, Table V and Figure 4 report the results for the 300-day training sets only; the results are similar for training sets of 300, 350 and 400 days.

The baseline model performs well, especially for countries with high frequencies of events, because these countries are likely to have different protests in consecutive days. The high recall increases the F1-score of the baseline for these countries. Nonetheless, in five of the six countries in Figure 4, the Lasso and hybrid models show better performance than the baseline model, in terms of higher mean and/or reduced variance of F1-scores. These improvements over the baseline are more pronounced for Argentina and Colombia. The boxplots for these countries, in addition to those for Brazil, Paraguay and Venezuela, demonstrate that the performance variation across the test sets is larger for baseline, which makes it less appealing than our models. Moreover, it does not give any information about the nature of social unrest. On the other hand, our methods provide insights about the underlying social dynamics in these countries by identifying features associated with civil unrest.

The features selected by the Lasso models as most relevant to civil unrest in the countries are shown in Table VI. As this analysis is exploratory and because accurate estimates of model uncertainty are not straightforward from Lasso, we do not report coefficient values and confidence intervals; instead, we focus on which features were selected for model inclusion based on their explanatory value in the model. The features

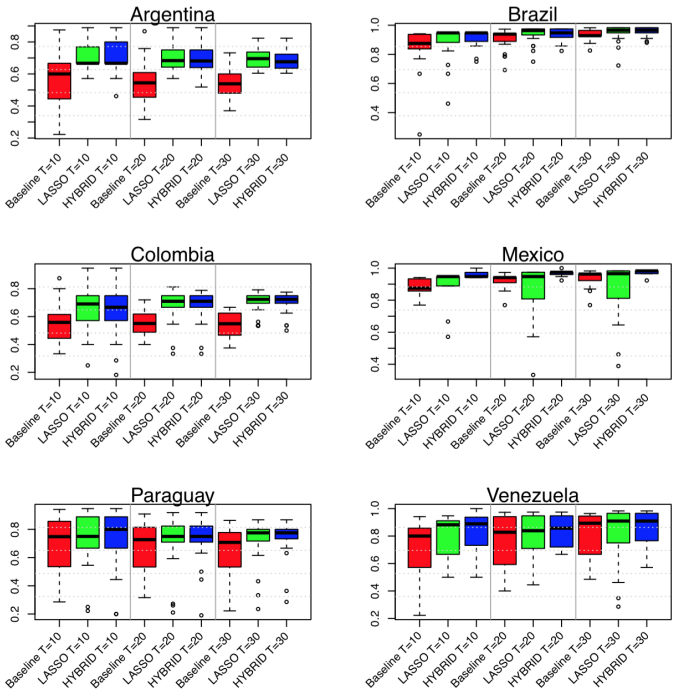

Fig. 4: Boxplots of the distributions of F1-scores for Baseline, Lasso and Hybrid models for 300 training days and testing periods of 10, 20 and 30 days, respectively.

TABLE VI: Top features selected by Lasso for unrest forecasting. The variables with the highest (top three) coefficients are highlighted in blue.

	NEWS	BLOGS	TWITTER	OTHER
Argentina	**SALDO BLANCO** **ÁNIMOS CALDEADOS** BOLSILLO	**COCALEROS** ADMINISTRACIÓN PÚBLICA GUERRA CIVIL	POLICÍA MILITAR GREMIOS DECLARAR	GDELT-19 (Fight)
Brazil	**COMERCIANTES** || URBANO **RACISMO** || MILITAR **ENCARAR** || COBRAR	PROTESTA OPERATIVO	DIRECTAMENTE PROYECTO	Currency
Colombia	**ASESINATO EXTRAJUDICIAL**	**OBRA HIDRÁULICA**	**RELACIÓN LABORAL** MANIFESTACIÓN PÚBLICA MINEROS INFORMALES	
Mexico	**ALCALDE**	**PROTESTA** **INCONFORMES**		
Paraguay	**RECHAZADOS**	**ABOGADO** **MOVIMIENTO** ELECCIÓN	ORGANIZADO	
Venezuela	**PROTESTA**	**DISTURBIOS** ALCALDE	**REPRESIÓN**	Tor

commonly selected by Lasso are often keywords from Twitter, news and blogs; the table reports the keywords with the highest positive coefficients. Lasso chooses the currency variable only for Brazil, and Tor is selected only for Venezuela among all the countries of this study. Regarding the features from political event databases, GDELT-19 (Fight) has a positive coefficient in the prediction model of Argentina.

As discussed in the next section, an analysis of the events in the GSR shows that various country-specific events are related to the variables selected for each country. Some examples are drug wars (Argentina, "cocaleros"), murders (Colombia, "asesinato extrajudicial"), and racism (Brazil, "racismo"). Different data sources can provide additional information about these events and the keywords picked by Lasso can be useful for interpretation of country-specific events.

C. Case Studies: Identifying Predictors of Protests

Here, we explore the variables selected by Lasso and the signals from different data sources for different types of events.

The case studies illustrate that multiple data sources are useful in understanding the context of the unrest events. Figure 5 illustrates the frequencies of the types of GSR events in Brazil and Venezuela. We observe that most of the unrest events that involve the general population in these countries fall under the "013 - Energy and Resources" and "Other Government Policies" categories. The 013 type of protests occurs due to the lack of availability or restrictions on use, or cost of anything that can be used directly as a source of energy. This includes gasoline, heating oil, natural gas, and electricity. This category also includes events that erupt due to the lack of materials or resources such as community resources (e.g., social services), natural resources (e.g. coal, oil, water, forests, and minerals), and health resources (e.g. services and materials provided for health and mental welfare). Loss of ownership, such as property or mineral rights, is included in this category. The 015 category includes events that occur due to government policies, mandates, regulations, etc. that negatively impact the population; pro-government demonstrations are included. For example, the October 2011 nationwide strike by students in Chile calling for education reform is a 015 - Other Government Policies event in the GSR. The majority of events in Venezuela fall under "015 - Other Government Policies" in the GSR, as do most events in Brazil. Event types 013 and 016 are also dominant in Brazil. We now address particular case studies.

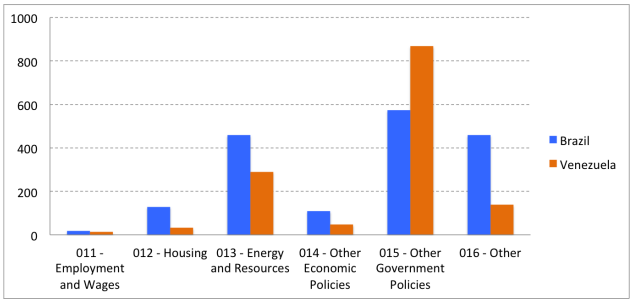

Fig. 5: Distribution of GSR Events in Brazil and Venezuela (November 1, 2012 - August 31, 2014).

1) Brazilian Spring, 2013: Our first case study is the chain of protests that took place in Brazil in June 2013, known as the Brazilian Spring. Social media was one of the main tools used for the coordination of these protests. The tipping point of the demonstrations was the increase of bus fare prices. The first big protest was held on June 6 on Paulista Avenue, one of the most important avenues of the Brazilian city of Sao Paulo. These uprisings involved the General Population and were categorized as "013 - Energy and Resources" events in the GSR.

The Lasso model for Brazil suggests that economic indicators, such as currency exchange rate and words related to economic activity, are valuable signals of civil unrest events. Two of the variables selected by our model are the currency indicator and the frequency of the word "comerciante" (Portuguese for trader or merchant). Figure 6 depicts the time series of the currency indicator for the duration of our study. There is a clear upward trend on currency coinciding with the start of the Brazilian Spring and subsequent variability in the months after the protests began. The figure shows that, during this period, there was an elevated number of days with civil unrest

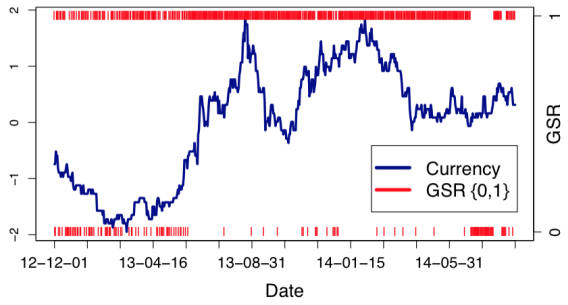

Fig. 6: The time series of the currency values (standardized), and the dates of GSR events in Brazil (1 if there is an event; 0 otherwise). The upward trend corresponds to the start of nationwide protests in Brazil in June 2013 (Brazilian Spring).

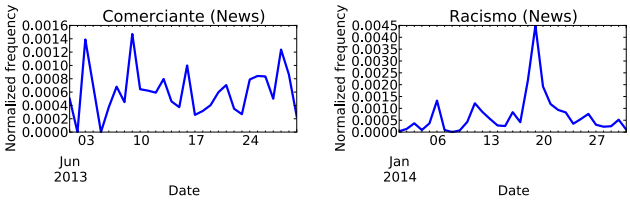

Fig. 7: The frequency of the keyword "comerciante" in news (left). Frequency of the keyword "racismo" in news (right). The peaks correspond to the GSR event dates in Figure 8.

events (as indicated by the red tick marks). Figure 7 (left) shows the time series of the keyword "comerciante" from news sources for June 2013. The peaks in the time series on June 3 and June 15 align with events in the GSR involving merchant groups in two cities of Brazil. The corresponding events in the GSR are shown in Figure 8, where the words highlighted in green correspond to variables selected by the Lasso model.

2) Racism in Brazil: Based on the Lasso-selected keyword "racismo", we show an additional example of the explanatory power of the keywords selected by our model. Figure 7 (right) shows time series of the word "racismo" (Portuguese for racism) in January 2014. There is an elevated use of this word in news media just a few days before a protest against racism took place in the capital city (Figure 8).

3) Venezuelan Protests, 2014: In 2014, Venezuela witnessed several nationwide protests. The main reasons were the indifference to student concerns, high levels of criminal violence, inflation, and chronic scarcity of basic goods due to strict price controls enforced by the government (see Figure 9). As shown in Figure 3, the time series of both Twitter and

Date	Description	Location
6/3/13	**Comerciantes** da Feira da Madrugada, no Brás, fizeram hoje (3) **manifestação** pelas ruas da capital paulista em **protesto** contra o fechamento temporário da feirinha pela prefeitura para realização de obras de segurança no local.	Sao Paulo, Brazil
6/17/13	Com gritos de ordem, faixas e cartazes de **protesto**, integrantes do Movimento Passe Livre começaram a **manifestação** no final da tarde desta segunda-feira, na região do Shopping Iguatemi, em Salvador, conforme combinado em reunião no último sábado. Eles seguiram em direção à Avenida Tancredo Neves, onde se localizam os principais escritórios **comerciais** da capital baiana.	Salvador, Brazil
1/25/14	Pouco mais de 30 jovens se reuniram hoje em frente ao Shopping Iguatemi - bairro nobre de Brasília, **protestando** contra forças de **racismo** e discriminação.	Brasilia, Brazil

Fig. 8: Examples of the civil unrest events in Brazil including their dates, descriptions, and locations in the GSR. The words highlighted in green are the keywords selected by the Lasso model from social media (see Table VI).

Fig. 9: Venezuelan protesters sign: "Why do I protest? Insecurity, scarcity, injustices, repression, deceit, for my future."

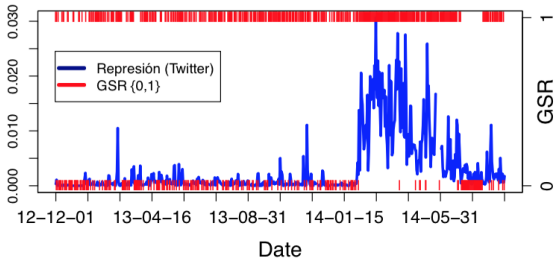

Fig. 10: The time series of the keyword "represión" in Twitter (normalized), and the dates of GSR events in Venezuela (1 if there is an event; 0 otherwise). The increased use of the word corresponds to the dates of student-led protests in Venezuela in February 2014.

blogs peak during the nationwide protests of February and March 2014. The keyword "represión" (Spanish for repression) in addition to Tor are picked by Lasso. Figure 10 illustrates that prior to and during the student-led protests in Venezuela in February 2014, there is a significant increase in the use of the keyword "represión" in Twitter messages. It is natural that the Lasso-picked variables are indicators of civil unrest given that the central government in Venezuela has strict measures against opposition demonstrations and control over most national media. In a country where expressing a negative opinion about the government could carry severe negative repercusions, anonimity tools, such as Tor, become important indicators of protest. Figure 5 illustrates that protests related to government policies constitute the most common protest type in Venezuela.

VI. Future Work

We are exploring the benefits of introducing other features, such as network measures based on interactions on social media. Future work is to construct a generalized dynamic model that incorporates changes in time.

Acknowledgements. This work has been partially supported by the following grants: DTRA Grant HDTRA1-11-1-0016, DTRA CNIMS Contract HDTRA1-11-D-0016-0010, NSF ICES CCF-1216000 and NSF NETSE Grant CNS-1011769. Also, supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] J. Lynch, *The Spanish-American Revolutions, 1808–1826*, 1973.

[2] F. F. Piven and R. A. Cloward, *Poor People's Movements*, 1977.

[3] M. F. Bellemare, "Rising food prices, food price volatility, and social unrest," *Am. J. of Agricultural Econ.*, vol. 97, no. 2, pp. 1–21, 2015.

[4] L. El-Katiri, B. Fattouh, and R. Mallinson, "The Arab uprisings and MENA political instability: Implications for oil & gas markets," Oxford Institute for Energy Studies, 2014.

[5] N. Ramakrishnan *et al.*, "Beating the news with EMBERS: Forecasting civil unrest using open source indicators," in *KDD*, 2014.

[6] F. Chen and D. B. Neill, "Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs," in *KDD*, 2014, pp. 1166–1175.

[7] S. Asur and B. A. Huberman, "Predicting the future with social media," in *WI-IAT*, 2010, pp. 492–499.

[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *ICWSM*, 2010, pp. 178–185.

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[10] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *1st Workshop on Social Media Analytics*, 2010.

[11] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter," in *ECML PKDD*, 2010, pp. 599–602.

[12] P. Chakraborty *et al.*, "Forecasting a moving target: Ensemble models for ili case count predictions," in *SDM*, 2014, pp. 262–270.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW*, 2010.

[14] M. Arias, A. Arratia, and R. Xuriguera, "Forecasting with twitter data," *ACM TIST*, vol. 5, no. 1, pp. 8:1–8:24, 2013.

[15] V. Wulf *et al.*, "Fighting against the wall: Social media use by political activists in a Palestinian village," in *SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 1979–1988.

[16] K. Starbird and L. Palen, "(how) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising," in *CSCW*, 2012.

[17] V. Wulf *et al.*, "'On the ground' in Sidi Bouzid: investigating social media use during the Tunisian revolution," in *CSCW*, 2013.

[18] N. Kallus, "Predicting crowd behavior with big public data," in *WWW Companion*, 2014, pp. 625–630.

[19] R. J. Stoll and D. Subramanian, "Hubs, authorities, and networks: Predicting conflict using events data," in *Annual Meeting of International Studies Association*, 2006.

[20] D. J. Gerner *et al.*, "Machine coding of event data using regional and international sources," *Inter. Studies Quarterly*, pp. 91–119, 1994.

[21] K. Leetaru and P. Schrodt, "GDELT: Global Data on Events, Language, and Tone, 1979-2012," in *International Studies Association Annual Conference*, 2013.

[22] M. D. Ward *et al.*, "Geographical models of crises: Evidence from ICEWS," in *2nd International Conference on Cross-Cultural Decision Making: Focus*, 2012, pp. 21–25.

[23] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan, "Detecting and forecasting domestic political crises: A graph-based approach," in *WebSci*, 2014, pp. 192–196.

[24] S. Muthiah *et al.*, "Planned protest modeling in news and social media," in *IAAI*, 2015.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Stat. Society. Series B (Meth.)*, pp. 267–288, 1996.

[26] L. Shi, N. Agarwal, A. Agrawal, R. Garg, and J. Spoelstra, "Predicting US primary elections with twitter," *URL: http://snap. stanford. edu/social2012/papers/shi. pdf*, 2012.

[27] J. Tang, X. Wang, and H. Liu, "Integrating social media data for community detection," in *Modeling and Mining Ubiquitous Social Media*, 2012, pp. 1–20.

[28] D. J. Gerner *et al.*, "Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions," *International Studies Association*, 2002.

[29] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, pp. 403–420, 1970.

[30] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.