Published in final edited form as:

Soc Netw Anal Min. 2016 ; 6: . doi:10.1007/s13278-016-0355-8.

Multi-Source Models for Civil Unrest Forecasting*

Gizem Korkmaz, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan

Abstract

Civil unrest events (protests, strikes, and “occupy” events) range from small, nonviolent protests that address specific issues to events that turn into large-scale riots. Detecting and forecasting these events is of key interest to social scientists and policy makers because they can lead to significant societal and cultural changes. We forecast civil unrest events in six countries in Latin America on a daily basis, from November 2012 through August 2014, using multiple data sources that capture social, political and economic contexts within which civil unrest occurs. 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 subject matter experts. We use logistic regression models with Lasso to select a sparse feature set from our diverse datasets. 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 method. The ablation study demonstrates the relative value of data sources for prediction. We find that social media and news are more informative than other data sources, including the political event databases, and enhance the prediction performance. However, social media increases the variation in the performance metrics.

1 Introduction

1.1 Background and Motivation

Civil unrest events (protests, strikes, and “occupy” 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 [18,22,3]. 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 [22]. Predicting the occurrence of civil unrest events is in the interest of policy makers, since local unrest can lead to regional instability [8].

* A preliminary version of the paper appeared in the Proceedings of 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining [15].

Gizem Korkmaz, Biocomplexity Institute of Virginia Tech, 900 N Glebe Rd, Arlington VA 22203, USA, Tel.: +1-571-858-3113, Fax: +1-571-858-3015, gkorkmaz@vbi.vt.edu.

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 2). 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

In this paper, we develop a statistical model to forecast civil unrest events in six Latin American countries using multiple data sources as shown in Figure 1. Specifically, we combine data from social media (Twitter, blogs, news), political event databases (ICEWS, GDELT), Tor statistics, and exchange rates as a proxy for economic condition. We use logistic regression with Lasso regularization to conduct experiments on a longitudinal data set of civil unrest events in Latin America over the course of 2 years, from November 2012 to August 2014.

1.2 Contributions

We present a model for predicting civil unrest through the combination of heterogeneous online data sources and provide a critical evaluation of the approach. To the best of our knowledge, this is the first model of social unrest forecasting that combines several relevant data sources, and explores the relative value of different data sources.² Our main contributions follow.

1. Forecasting of civil unrest events using social, economic and political indicators—We develop statistical models (Lasso, Group-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 websites, 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 those of the overall

¹June 21, 2013, “Protesters, criminals get around government censors using secret web network,” <http://bit.ly/1Sghvo7>.

²This model was briefly mentioned, along with several others in [23] as part of an automated, real-time forecasting software system. This paper describes our model and results in detail.

system [23] 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 5.2.

2. Interpretable feature selection—An important attribute for predictive models of the type we seek here is that the model should down-select the variables, from a potentially large candidate set, that yield good predictions, Lasso-based models do this. As an example, we have 2,988 features on which our models are based. Even when predictions use training sets that span years, with daily outcomes for training, the number of features, $p = 2,988$, may be much greater than the number of observations, $n = 669$ days (~2 years). Analyses of thousands of civil unrest events herein indicate that the number of features that Lasso selects is only a small fraction of the total feature set; ranging from 2 up to 119 (see Table 10). We show that the features that Lasso models extract give insights into the nature of civil unrest events and have the potential to provide actionable information to policy planners.

3. Evaluation across different Lasso-based models and comparisons with other models—We critically compare Lasso, Group-Lasso, and a hybrid model based on F1-scores, and precision and recall. For instance, we explain why the hybrid model does not produce significantly better results than the Lasso model. These comparisons are made across different training periods, from 300 to 400 days, across different test periods of between 10 and 30 days, and across six countries. Using Lasso-identified features, we also use a logistic regression model (designated GLM) over the entire dataset, and we compare this model's predictions with those from Lasso model. (Note that this is another benefit of Lasso's ability to select relevant features which then can be used in other models.) We also compare our models against a baseline model that predicts the occurrence of an event on a particular day, based on whether an event occurred on the previous day. Since several of the countries we evaluate have many civil unrest events, the baseline model does a decent job of predicting social unrest. However, the variance in the predictions is greater than that of the Lasso model for some countries, and it does not provide any insights into the events themselves because there is no feature selection (see Section 5).

4. 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 7. 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 (to avoid 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 selected by Lasso provide additional information about various unrest events. Moreover, the case studies span multiple countries, multiple cities within a country, and multiple dates of events.

5. Value of combining 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 3.1). 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, 23]. In Figure 4, 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. 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 7). The third finding is surprising in that political event databases do not provide particularly strong signals of protests and civil unrest. This is surprising because unrest events, according to the GSR, predominantly stem from government policies (Figure 9).

6. Evaluation of the importance of data sources through ablation studies—

Ablation studies—purposely omitting combinations of data sources to determine the resulting performance of the prediction model—were performed over all six countries. For each country, seven different combinations of sources were evaluated (including the model that uses all data sources). These use receiver operating characteristic (ROC) curves as a basis for comparison. This study shows that social media is more informative than other data sources, and enhances the model fit (significantly increasing the area under the ROC curve, AUC). However, social media also increases the variance in AUC. Surprisingly, but consistent with the observation above, political event databases are not as informative as social media, but the ROC curves generated based on political event databases do not vary much compared to the social media (see Section 8).

7. Extensibility of methods to other types of events—The Lasso model and ablation studies indicate that the most important variables for predicting the social unrest events that are evaluated in this work come from news sources, blogs, and social media. The model uses daily volumes (i.e., counts) of keywords that appear in these sources. (The keywords were compiled by a panel of domain experts; see Section 3.) The model is agnostic to these keywords. The implication is, therefore, that a model such as the one developed here might be useful for predicting other types of events. For example, policy planners and disaster managers are interested in using social media to gauge public sentiment and (planned) actions; e.g., how many people might be evacuating a coastal region in the face of an impending hurricane, and how these numbers are predicted to change daily. With Lasso models, the keywords could be changed to a set that reflects this type of event, with the underlying model remaining the same. Clearly, this has not been demonstrated, and other aspects of model building such as training the model would have to be performed, but there are general aspects of the model that might be exploited for other domains.

2 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 [2], political elections [32], the stock market [4], flu trends [7,16,5], and even earthquakes [24]. A summary of the different predictive tasks studied and the proposed methods can be found in [1]. More recently, Twitter has also been used to forecast civil unrest [6, 23]; however, most of the previous work involving Twitter has focused on how people interact on Twitter in times of protest [34,26,35] 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. Kallus [13] 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 [27] 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 [9] and GDELT [17]. 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 [33, 14].

Two additional data sources are considered here: Tor usage metrics and currency exchange rates. These data sources have not been used as predictors of civil unrest or related topics in the literature. Furthermore, 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 [5], 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 [23] was the fusion of predictions from multiple *models* for real-time forecasting of flu. One of those models [21] uses multiple sources to identify protests that are called out in different media, however the data sources are not combined in a model.

Lasso regression [31] has garnered interest in diverse fields in recent years. One appealing property of this technique is that irrelevant features given to the model can be penalized and discarded by the regression, which encourages sparse representation and interpretability. In social media analytics, specifically, Lasso has been used to train models for predicting elections [25] and detecting flu with Twitter data [16]. One application of Lasso in social media with a similar goal to ours (i.e. combining different data sources) can be found in [29], where the authors integrate data from the social media sites BlogCatalog and Flickr for community detection.

3 Data Sources

3.1 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 drivers of the unrest such as “Wages and Employment” and “Energy and Resources” 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 1 shows a partial entry from the GSR about a nationwide protest that took place on 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 2 illustrates the frequency of events in Latin America. In this paper, we focus on the six countries with the highest volume of civil unrest events (Argentina, Brazil, Mexico, Colombia, Paraguay and Venezuela).

3.2 Data Sources for Forecasting Models

Social media (Twitter, blogs) and news—Social media has played an important role in the organization of public outcries [12] and has served as a primary communication platform for protesters in recent civil unrest demonstrations [30]. This 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 2 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 on social media is summarized by filtering and aggregating our datasets using a dictionary of 962 protest-related keywords that was compiled by political scientists and domain experts on Latin America.⁴ The dictionary, which is illustrated partially as a randomly-sized word cloud in Figure 3, 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

³<http://lanic.utexas.edu/>

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

“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 4. 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. Moreover, the peaks in one social media channel in absence of an event could also be suppressed by using multiple data sources.

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 have been used for the prediction of large-scale demonstrations and political crises (see Section 2). Two such datasets are considered in our study: Integrated Crisis Early Warning System (ICEWS) [9] and Global Data on Events, Location and Tone (GDELT) [17]. Both of these databases classify events using a CAMEO (Conflict and Mediation Event Observations) framework, which is an event data coding scheme optimized for the study of third party mediation in international disputes [10]. 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 3 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 from ICEWS and GDELT 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—Authoritarian regimes have been trying to prevent the communication among protesters by controlling social media. During the nationwide Venezuelan protests of

February 2014, the government blocked users' online images as opposition groups marched through Caracas.⁵ Many countries have experienced blocks on these channels, especially on Twitter and YouTube, as well as arrests due to online posts against the government. A Twitter user recently got arrested for insulting the president of Turkey by referring to him as a "dictator."⁶ The Onion Router⁷ (Tor) is a free anonymization tool that protects an individual's identity and location on the Internet. The Tor daily usage statistics are used as 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. Currency, by itself, cannot explain the economic stability of a country. However, it could be considered as an additional indicator of the instability and discontent in a country.

4 Prediction Models

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 $\mathbf{X}_t \in \mathbb{R}^{p+1}$ be a *feature vector* (including the constant term) that summarizes the data corresponding to a day $t \in \{1, 2, \dots, n\}$, and let $y_{t+1} \in \{0, 1\}$ be an indicator variable that equals 1 if there is a civil unrest event on day $t + 1$. The goal is to find a function $f: \mathbb{R}^{p+1} \rightarrow \{0, 1\}$, such that

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

At the same time, we are interested in finding which features are good predictors of protest in the countries under 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 4). 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.

4.1 Lasso Model (Logistic regression with Lasso regularization)

For our prediction task, we develop a logistic regression model since we have a binary dependent variable. Let $\mathbf{Y} \in \{0, 1\}^n$ represent the binary vector of response variables, and $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ be a *feature matrix* representing the various data sources discussed in Section 3.

⁵Source: Bloomberg, <http://www.bloomberg.com/news/articles/2014-02-14/twitter-says-venezuela-blocks-its-images-amid-protest-crackdown>

⁶Source: Yahoo news, <http://news.yahoo.com/turkey-arrests-3-raids-over-erdogan-twitter-insults-155612123.html>

⁷Source: <https://www.torproject.org/>

Each row, $\mathbf{X}_t = (1, X_1, X_2, \dots, X_p)$, is a feature vector for day $t \in \{1, 2, \dots, n\}$. Formally, we estimate

$$\log \left(\frac{\Pr(y_{t+1} = 1)}{1 - \Pr(y_{t+1} = 1)} \right) = \mathbf{X}_t \beta \quad (2)$$

where $\beta \in \mathbb{R}^{p+1}$ is the vector of coefficients to be estimated (p features and one intercept term), and $y_{t+1} \in \{0, 1\}$ 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.

Since the features of our model outnumber observations ($p \gg n$), we employ Least Absolute Shrinkage and Selection Operator (Lasso) regression for feature selection [31]. 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. The objective function for the penalized logistic regression uses the negative binomial log-likelihood, and Lasso computes a sparse regression estimate vector $\hat{\beta}$, by solving the following optimization problem:

$$\min_{\beta \in \mathbb{R}^{p+1}} - \left[\frac{1}{n-1} \sum_{t=1}^{n-1} y_{t+1} \cdot \mathbf{X}_t \beta - \log \left(1 + e^{\mathbf{X}_t \beta} \right) \right] + \lambda \|\beta\|_1 \quad (3)$$

where $\lambda \geq 0$ 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 Lasso method can perform variable selection. The regularization parameter is determined by cross validation. Finally, the estimated probability of an event on day $t + 1$ given the set of features on day t can be computed as

$$\hat{P}_{t+1} = \frac{e^{\mathbf{X}_t \hat{\beta}}}{1 + e^{\mathbf{X}_t \hat{\beta}}} \quad (4)$$

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 [11]. 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 also are interpretable.

4.2 Group-Lasso Model (Logistic regression with Group-Lasso penalty)

There is a group analog to Lasso, called *Group-Lasso* [36], that shrinks entire groups of coefficients to zero. Suppose there are G groups of variables (with different sizes), and let the feature matrix for group i be denoted by $\mathbf{X}_{t,i}$. In this case, we estimate

$$\log \left(\frac{\Pr(y_{t+1} = 1)}{1 - \Pr(y_{t+1} = 1)} \right) = \sum_{i=1}^G \mathbf{X}_{t,i} \beta_i. \quad (5)$$

Let $\pi(X_t)$ denote $\sum_{i=1}^G \mathbf{X}_{t,i} \beta_i$. The logistic Group-Lasso obtains the estimates $\hat{\beta}_i$, $i = 1, 2, \dots, G$, as the minimizer of the objective function:

$$-\left[\sum_{t=1}^{n-1} y_{t+1} \cdot \pi(X_t) - \log(1 + e^{\pi(X_t)}) \right] + \lambda \sum_{i=1}^G \gamma_i \|\beta_i\|_2. \quad (6)$$

The parameter λ controls the amount of regularization, with larger values implying more regularization. When λ is large enough, all the coefficients will be estimated as zero. The γ_i 's allow each group to be penalized to different extents, which allows us to penalize some groups more (or less) than others. If each group consists of only one variable, this reduces to the previous Lasso criterion. Just as Lasso performs variable selection by estimating some of the coefficients to be zero, the Group-Lasso does selection on the group level. It is able to zero out groups of coefficients. If an estimate $\hat{\beta}_i$ is nonzero, then all its components are usually nonzero.

Whether or not adding extra constraints on groups of variables is not clear *a priori*. On one hand, adding information about groups allows the model to discard an entire group of irrelevant predictors. It is also more intuitive to look at the output and see which groups of variables are important. On the other hand, by adding the group constraint, we may force Group-Lasso to take in some non-relevant predictors. It is *all-or-nothing*: We either take or discard the entire group. We evaluate two natural ways of grouping features:

- i. **Source-based.** In this model, variables are grouped based on the data sources; thus, we use 7 groups: Twitter, news, blogs, Tor, currency, ICEWS, and GDELT. Grouping the predictors by source allows the model to find which data sources are relevant for the prediction task.
- ii. **Keyword-based.** In this version of Group-lasso, 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. Grouping by keyword allows the model to find which keywords are important for prediction regardless of the data source where the keyword was originated. As we see in Section 5.2, these two models have different behaviors; however, neither of the two increases the predictive performance (defined below) compared to the regular Lasso.

4.3 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

$$\hat{y}_{t+1} = \begin{cases} 0 & \text{if } y_t = 0 \\ 1 & \text{if } y_t = 1 \end{cases} \quad (7)$$

This particular baseline model is chosen based on the fact that the countries of this study have many protests (see Fig. 2 and Table 5). In order to more stringently evaluate this multi-source model, we use a baseline model that predicts many protests. This model sets a benchmark in order to measure the added predictability provided by the multiple datasets.

4.4 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:

$$\log \left(\frac{Pr(y_{t+1} = 1)}{1 - Pr(y_{t+1} = 1)} \right) = \mathbf{X}_t \beta + \theta y_t. \quad (8)$$

The hybrid model uses a Lasso logistic regression for feature selection, but it also imposes a mandatory binary variable, y_t , with its coefficient θ , that is independent of the Lasso selection. Therefore, the optimization problem is similar to the one in Eqn. 3 with the additional θy_t term.

4.5 GLM Model (Logistic regression with Lasso-selected features)

This model is used in order to show the goodness-of-fit of a logistic regression with features selected by Lasso because accurate estimates of model uncertainty are not straightforward from Lasso. As a first step, Lasso is executed for variable-selection as in Eqn. 2 since the total number of features outnumbers the observations. The subset of features that are selected by Lasso are used in a second logistic regression without Lasso penalty, λ . Here, we estimate a standard binomial logistic regression model

$$\log \left(\frac{Pr(y_{t+1} = 1)}{1 - Pr(y_{t+1} = 1)} \right) = \tilde{\mathbf{X}}_t \tilde{\beta} \quad (9)$$

where $\tilde{\mathbf{X}}_t$ is the vector of features that are selected by Lasso, i.e., the features with non-zero coefficients, $\hat{\beta} = 0$, from Eqn. 3. The maximum likelihood estimates are obtained using the standard log likelihood function:

$$\sum_{t=1}^{n-1} y_{t+1} \cdot \tilde{\mathbf{X}}_t \tilde{\beta} - \log \left(1 + e^{\tilde{\mathbf{X}}_t \tilde{\beta}} \right). \quad (10)$$

This allows us to compute the p-values and the R^2 , and explore the fit of the model that use the Lasso-selected variables as independent variables. This model and the results are discussed in detail in Section 6.

5 Experiments and Prediction Results

The experiments in this section are designed (i) to assess the predictive ability of our models, (ii) to evaluate the robustness of the predictions of our models for different training periods, and over different periods of time across the whole dataset, (iii) to compare the predictive performance of our models against the baseline in terms of precision, recall and F1-score (defined below), and (iv) to identify the relevant features in capturing events in the countries of interest. We begin by describing our evaluation methodology and performance metrics.

5.1 Experimental Methodology

Preprocessing—Before training the models, every variable is standardized to have mean zero and unit variance. We generate the feature vector $\mathbf{X}_t$ by combining the variables from all of the data sources. Due to missing data as a result of issues during the collection of social media data sources (Twitter, news, and blogs), we remove the rows of the combined feature matrix corresponding to those days. In other words, entire day must be removed if any of the data sources does not exist on that day. We also remove the respective rows from the GSR, i.e., y_{t+1} . Therefore, we obtain different numbers of observations for different countries. Table 5 summarizes the number of observations for each country and the number of days with at least one GSR event.

Metrics—The performance of our prediction models is evaluated using standard classification metrics:

- Recall (True Positive Rate): the percentage of GSR events that the model correctly predicts. That is, what percentage of events is the model catching?
- Precision: the percentage of event predictions of the model that are actually matched with a GSR event. That is, when the model suggests there will be a protest, what percentage of the time is there actually a protest?
- F1-score: the weighted harmonic average of precision and recall. A model that predicts an event every day, i.e., $\hat{y}_{t+1} = 1 \forall t$, will have a recall of 1, whereas a model with very few positive predictions will have a high precision. The F1-score is used to control for these biases.

Training and testing—In order to assess the robustness of the models, different training periods of 300, 350, and 400 days are used. 10-fold cross validation approach is used on the training sets to tune the regularization parameter λ . The cross validation partitions the data into 10 equally sized segments. One fold is held out for validation while the remaining folds are used to train the model and then used to predict the response variable in the hold-out set. This process is repeated 10 times. 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 :

$$\hat{y}_{t+1} = \begin{cases} 0 & \text{if } \hat{P}_{t+1} \leq \tau^* \\ 1 & \text{if } \hat{P}_{t+1} > \tau^* \end{cases} \quad (11)$$

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 ten-fold cross validation approach results in different feature selection and performances for different executions of Lasso. The model with the highest F1-score is chosen over multiple executions of Lasso for each training and test periods. 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.

5.2 Performance Results

Table 6 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); the average values over the multiple testing sets for 300-day training period 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 to 0.91. Precision is lower for Argentina and Colombia, which are the countries with fewer numbers of events.

In Table 6, we observe that the average F1-scores are in the range 0.68 to 0.95 and their distributions are illustrated in Figure 5. Each box plot 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 medians of the F1-scores. Table 6 and Figure 5 report the results for the 300-day training periods; the results are similar for training sets of 300, 350 and 400 days.

The baseline model, illustrated as red box plots in Figure 5, 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 5, the Lasso and hybrid models show better performance than the baseline model, in terms of higher median and/or reduced variance of F1-scores. These improvements over the baseline are more pronounced for Argentina and Colombia. The box plots 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 Group-Lasso does not improve the F1-score compared to the regular Lasso model. In the source-based model (Section 4.2), where the features are grouped by data source, entire social media sources —like Twitter or news— are discarded by the model. An explanation for this is that most keyword features coming from these data sources are not relevant for the prediction task. Given the group constraint, it is better —from an optimization point of view— to discard the entire group, thus losing valuable predictors. When we trained the model, we encountered convergence problems, thus we omit these results. The keyword-based model, which groups the counts of the same keyword in each text-based data source (Twitter, news and blogs) and takes the features from other data sources individually, has 1,064 groups. Table 7 summarizes the results for Brazil, Mexico and Venezuela. The remaining countries perform poorly with this model compared to the regular Lasso. Interestingly, we observe that the Group-Lasso increases the precision and lowers the recall for these countries. The model finds a *focused set* of keywords that are related to civil unrest regardless of the data source. However, the overall effect on the F1-score is not positive due to the loss in recall.

The features selected by Lasso as most relevant to civil unrest in the countries are shown in Table 8. We focus on which features were selected for model inclusion based on their explanatory value in the model; we report the selected features that result in the highest F1-scores. The features commonly selected by Lasso over the training sets 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 selected by Lasso can be useful for interpretation of country-specific events.

6 Model Fit and Significance Analysis

Our results from Section 5 show that a logistic regression with regularization has good prediction power for our civil unrest setting. We have also demonstrated the robustness for different training and test periods. In this section, our goal is to determine the model fits using summary measures of goodness-of-fit over the entire data. We perform a model fit and significance analysis to understand why Lasso model shows good prediction performance. In this model, we use the entire data set since our goal is not prediction, but to explore goodness-of-fit.

As a first step, we execute Lasso for feature selection (since $p \gg n$) over the entire dataset, and then use the features selected by Lasso (i.e., those with non-zero coefficients) to fit a second logistic regression (referred to as “GLM model”, described in Section 4.5). We use 10-fold cross validation for the logistic regression with Lasso penalty. As described in Section 5.1, this results in different feature sets and performance levels for different executions of Lasso. Since our goal is to explore the model fit, and not to focus on prediction performance, here we do not choose the model with the highest F1-score. Instead, we aim to capture the variation in the model fits by executing Lasso model 50 times, and for each selected feature set we execute the GLM model, i.e., fit a second logistic regression without a Lasso penalty. In this section, we report the goodness-of-fit of this model (GLM) for the 50 executions of Lasso.

Metrics

The fit of the model is evaluated using the following metrics:

- Deviance: is a measure of goodness-of-fit of a generalized linear model. The *null deviance* shows how well the response variable is predicted by a model that includes only the intercept; the *residual deviance* shows the reduction in the null deviance with the inclusion of independent variables. In order to test for significance, a chi-square test is conducted and p-values are computed.
- McFadden’s R^2 [19]: one of the standard measures for logistic regression that inherits the properties of the familiar R^2 from linear regression, defined as

$$R^2_{McFadden} = 1 - \frac{\log(L_c)}{\log(L_{null})} \quad (12)$$

where L_c denotes the (maximized) likelihood value from the current fitted model, and L_{null} denotes the corresponding value but for the null model - the model with only an intercept and no covariates. The measure ranges from 0 to just under 1, with values closer to zero indicating that the model does not fit the data well.

ROC curve (Receiver Operating Characteristic) and the area under the ROC curve (AUC): The curve is created by plotting the recall against the false positive rate (FPR) at various discrimination thresholds. AUC, the area between the ROC curve and the

no-discrimination line, is equal to the probability that the model will rank a randomly chosen positive instance (i.e., GSR event) higher than a randomly chosen negative one. Since we do not train the models to determine the threshold τ^* as in Eqn. 11 in Section 5, here we use the ROC curve, which is plotted for all possible thresholds, and AUC is used as an alternative measure for goodness-of-fit. This metric ranges from 0.50 to 1.00, and values above 0.80 indicate that the model does a good job in discriminating between the two categories, which comprise our target variable.

Table 9 reports the summary of the goodness-of-fit for the GLM model. We report the median values of null and residual deviances, null and residual degrees of freedom (DF), McFadden's R^2 , and AUC, respectively, over the 50 executions. The *null degree of freedom* equals the number of observations minus one (the intercept), and it is different for each country due to missing data (illustrated in Table 5). The *residual degree of freedom* is the difference between number of observations and the number of Lasso-selected variables used in the GLM model. Based on the reduction in the null deviances, we observe that for all countries, there is a significant improvement with Lasso-selected variables compared to a null model that just considers the intercept of the regression. We also use the residual deviances to test whether the null hypothesis is true (i.e., logistic regression model provides an adequate fit for the data) for each country. This is possible because the deviance is given by the chi-squared value at a certain degree of freedom. In order to test for significance, we compute associated p-values using the values of residual deviance and DF, we get a p-value of approximately zero for all cases showing that there is a significant lack of evidence to support the null hypothesis. A chi-square test for goodness-of-fit also reveals that most of the variables selected by Lasso are significant at the 5% level.

McFadden's R^2 can be interpreted as the standard R^2 such that the larger values are better than smaller ones, but McFadden's R^2 's are not expected to be as big as R^2 . A rule of thumb is that values between 0.2–0.4 indicate a very good model fit [20]. Table 9 illustrates that the median values for countries such as Argentina, Brazil, Mexico and Venezuela are between 0.24 and 0.54. However, the GLM models for Colombia and Paraguay result in a low R^2 . The number of features used in the GLM model for Colombia is found as 4 (convovar-Twitter, excepcional-Twitter, caminata-blogs, conflicto-blogs); and the only feature selected by Lasso for Paraguay is llamar-blogs. We observe that small number of features results in lower performance.

Table 10 illustrates the summary statistics of $R^2_{McFadden}$, AUC, and the number of features selected by Lasso. We observe that the number of features is positively correlated with performance for all the countries. For Paraguay, the model does not select many variables; it ranges from 1 to 10. The highest GLM performance is obtained when Lasso selects 10 variables; in this case an R^2 of 0.11 and an AUC of 0.71 is reached. Colombia, on the other hand, has a low median but the maximum R^2 can reach to 1 when the number of variables selected by Lasso is 119. The variation in the performance is very high for this country, ranging from 0.06 to 1.00. For the remaining countries, we observe a high performance measured by $R^2_{McFadden}$ and AUC, which increases with the number of features.

In Figure 6, we illustrate the ROC curves for the GLM model corresponding to the particular instances (the median AUC) over 50 executions. Except for Colombia and Paraguay, the AUC is above 0.80, indicating that our predictors fit the response variable well for these countries.

For the same instances, we examine performance in an alternative way in order to gain more insights regarding the performance. A property of a good model should be that when there is no event, the probability predicted by the model should be lower than the probability predicted when there is an event. This way, one can use a threshold and improve predictions. We show that this is the case with the logistic regression model. The box plots in Figure 7 illustrate the distribution of predicted values (i.e., estimated probabilities, $\hat{Y} \equiv \hat{P}_{t+1}$ as in Eqn. 11), when there is an event, $Y \equiv y_{t+1} = 1$, and when there is not an event, $Y = 0$. We observe that for most of the countries, there is a big separation between the distribution of the predicted probabilities for $Y = 0$ and $Y = 1$. When there is an event in the ground truth, the fitted probabilities are high and close to 1, and vice versa when the ground truth is zero. This shows that the model is a good discriminant of days with events and days without events. For example, a threshold of 0.5 would do very well at the prediction task for Argentina. We observe that the box plots for $Y = 1$ are significantly higher for all countries except for Colombia and Paraguay. This again explains why the Lasso model has a good performance for prediction for Argentina, Brazil, Mexico and Venezuela; however, the predictive capability is lower for Colombia and Paraguay.

Finally, the entire values of AUC (not only for particular instance) over 50 executions and the variation for each country are illustrated using box plots in Figure 8. Here we also illustrate the ROC curves for the logistic regression with Lasso penalty (Lasso model with the entire dataset) as well as the GLM model (logistic regression with Lasso-selected features) for each execution. We observe that the ROC curves obtained with Lasso always have lower AUC compared to the GLM with the corresponding Lasso-selected variables. Figure 8 illustrates the distribution of the AUCs for Lasso and GLM with Lasso-selected features. The box plots of GLM are significantly higher than those for Lasso, especially for Argentina, Brazil and Mexico. In Figure 8, we can observe the variation in the AUCs in addition to the particular values illustrated in Figure 6. The AUC for Colombia with both Lasso and GLM can reach up to 1, however the high variation suggests that our model for this country is not as consistent as for other countries.

7 Case Studies: Identifying Predictors of Protests

This section aims to provide insights about the social dynamics in the countries of interest using case studies. 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 unrest events.

We focus on Brazil, Mexico and Venezuela due to the richness of events in these countries. Figure 9 illustrates the frequencies of the types of GSR events in Brazil, Mexico 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, nationwide strikes by students calling for education reform are coded as 015 - Other Government Policies in the GSR. The majority of events in Brazil, Mexico and Venezuela fall under “015 - Other Government Policies.” The events that do not fall under any of these categories are codes as “016 - Other.” Event types 013 and 016 are also dominant in Brazil and Mexico.

The features selected by Lasso for Brazil, Mexico and Venezuela are illustrated as word clouds in Figure 10. The size of the features in these word clouds reflects the frequency of being selected by Lasso over all executions as described in Section 6. The largest ones are selected in all 50 instances. The number of features selected in all 50 executions is much higher for Brazil compared to Mexico and Venezuela, resulting in a larger word cloud for Brazil. The summary statistics of the number of selected features were illustrated in Table 10. The average number of selected features over 50 instances is 42.2 for Brazil, whereas it is almost half for Mexico and Venezuela).

We now address particular case studies for these countries. Table 11 summarizes these events and the relevant feature categories, along with the most frequently used keyword from news, blogs and Twitter. We can also identify these features in the respective word clouds in Figure 10.

7.1 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 the value of the currency and words related to economic activity, are valuable signals of civil unrest events. Currency, is used interchangeably with the exchange rate, and is defined as the value of the currency corresponding to 1 US Dollar (a higher value implies depreciation, that is the loss of value of a country’s currency). Two of the variables selected by our model are the currency variable and the frequency of the word “comerciante” (means “trader” or “merchant” in Portuguese). Figure 11 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 is an increase in the number of days with civil unrest events (as indicated by the red tick marks). Figure 12 (left) shows the time series of the keyword “comerciante” in news sources during June 2013. The peaks in the time series align with events in the GSR involving merchant groups in two cities of Brazil. The corresponding events in the GSR are shown in Figure 13, where the words highlighted in green correspond to variables selected by Lasso. This case study illustrates that our methodology allows for analysis of events across multiple cities and dates.

7.2 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 12 (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 13).

7.3 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 14). As shown in Figure 4, the time series of both Twitter and blogs peak during the nationwide protests of February and March 2014. The keyword “represión” (Spanish for repression) in addition to Tor are selected as predictors by Lasso. Figure 15 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-selected 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 repercussions, anonymity tools, such as Tor, become important indicators of protest. Figure 9 illustrates that protests related to government policies constitute the most common protest type in Venezuela.

7.4 Mexico: “Protesta” in Twitter, news and blogs

The predictors selected by Lasso for Mexico are, as in the previous two cases, related to events specific to this country as shown in Figure 10. A notable predictor is “privatizar la educacion” (privatizing public education), which is consistent with the fact that a large number of the protests in this country are related to education policies.

A broader predictor is the keyword “protesta” from the blogs dataset. It is interesting that the model finds the signal from blogs to be a better predictor than the signal from Twitter or news sources. To understand why this is the case, we compare the time series of GSR events in Mexico to the time series of the word “protesta” in three of our datasets (Twitter, news and blogs). Figure 16 illustrates the usage of the keyword in blogs (shown in purple – the chart on the right) is more reflective of the trends of events in the GSR than either Twitter (left) or news (center). Particularly, the time series of blogs better capture the rise in the number of events in the GSR from January 2014 to June 2014, whereas the level of activity

on Twitter for this period is the same as in 2013. In July 2014, the number of GSR events decreases abruptly, and the blogs' time series stay aligned with this trend. The news-based time series show decreased activity in July 2014 but the trend is not as pronounced as it is for blogs. We measure correlation formally by computing the distance correlation [28] between the GSR time series and the three data sources. The GSR has a distance correlation of 0.22 with blogs compared to 0.19 with news and 0.07 with Twitter (almost no correlation), so "protesta" in blogs is indeed a better indicator of GSR events.

7.5 Mexico - July, 2014

Instead of focusing on a particular feature or a specific event in Mexico, here we aim to compare the total volume of activity in Twitter and blogs against the GSR. Figure 17 illustrates a significant reduction of GSR events (in dark green) in July, 2014 in Mexico. The total volume of keywords in Twitter (top-left) does not seem to decline during the same period. However, the volume of keywords in blogs (top-right) follows the same trend as that of the GSR in July, 2014. This suggests that in absence of blogs, the model is likely to generate false positives due to the high volume of activity on Twitter. The blue chart in the bottom illustrates the time series of the estimated probabilities, $\hat{P}_{t+1}$ in Eqn. 11, generated by our model, which combines all data sources. The reduction in the probability in July, 2014 is evident in the figure. This supports our claim that combining data sources provides additional information that might lead to increase in true positives as well as decline in false positives.

8 An Ablation Study

In this section, we explore the relative value of each data source for our prediction task by conducting an ablation study – purposely omitting combinations of data sources to determine the resulting performance of the prediction model. We analyze the performance of the model for civil unrest prediction using seven different combinations of data sources:

- i. All (Twitter, news, blogs, ICEWS, GDELT, Tor, currency),
- ii. Social media (Twitter, blogs, and news),
- iii. ICEWS, GDELT, Tor, and currency,
- iv. Political event databases (ICEWS and GDELT),
- v. Tor and currency,
- vi. ICEWS,
- vii. GDELT.

The first model uses all data sources described in Section 3. The second combination includes our text-based data sources (Twitter, news and blogs), and this model is referred to as social media in this section. In the third one, we remove the social media data sources and keep the remaining ones to observe their effect to the overall model. This allows us to explore the value of Twitter, news and blogs for predicting civil unrest. The fourth combination uses the political event databases, and the fifth one includes Tor and currency,

which are the single-value data sources. We also use ICEWS and GDELT separately in order to compare these political event databases.

For each combination of data sources, we execute Lasso for variable selection over the entire dataset using 10-fold cross validation as in Section 6. We obtain ROC curves using the estimated probabilities and the GSR, and compare the area under the ROC curve (AUC) for each combination. Since Lasso selects different combination of variables with 10-fold cross validation, we explore the results of each 50 executions of Lasso. The box plots in Figure 18 illustrate the range of AUC values for all executions under each different combination of data sources. The ROC curves corresponding to the median AUC are given in Figure 19. Finally, Figure 20 illustrates the ROC curves for all executions of Lasso under different combinations of data sources, and the median AUC values over the 50 instances are given in parenthesis. This figure (conveys similar information as in Figure 18) is included to stress the high variation with social media data sources; the high range of ROC curves can be seen especially for Colombia.

Based on these figures, we observe that the models with all data sources and social media lead to higher values of AUC compared to other combinations. The performance of the social media is followed by models that include both of the political event databases. The models, which are referred to as IcwGdl and IGTC in Figure 18, behave similarly for most of the countries because Tor and currency do not provide much additional value. The model with these two data sources (TorCurr) result in the lowest performance.

8.1 The Power of Social Media (Twitter, blogs, and news)

The models with social media data sources (Twitter, news, and blogs) result in AUC values with the highest median compared to other combinations (including the all-data model) as illustrated in Figures 18 and 19. Moreover, Figure 20 demonstrates that the ROC curves of social media and all data models are very similar and aligned in most of the executions.

The removal of social media data sources result in a reduction in the median of the AUC values for all countries, and this reduction is significant for some. Especially for Argentina, Figure 20 illustrates the clear separation of the ROC curves of the social media model with the rest of the data sources. The best performance in absence of social media (IGTC) is achieved for Venezuela compared to other countries, the AUC is in the 0.75–0.77 range.

We can conclude that social media improves the prediction performance for all countries. Figure 20 illustrates that the curves in blue always reach higher compared to the other ones. The values of AUC for Brazil and Colombia with only social media reach towards 0.94 and 0.96, respectively. On the other hand, social media results in a much wider range of AUC values compared to other models. We can observe the high variation of ROC curves for these countries in Figures 18 and 20. Especially for Colombia, the area is ranging from 0.61 to 0.96; in some cases the AUC is even lower than the ones obtained under models with political event databases.

8.2 ICEWS vs. GDELT

The features from political event databases, ICEWS and GDELT, are not selected as frequently as variables from Twitter, news and blogs. In the absence of social media, however, we observe that both databases turn out to be informative, and the models result in AUC values with a very low variation. The IcwGdl model for Brazil results in the largest AUC compared to other countries, followed by Mexico and Venezuela.

When we compare the political event databases, we observe that the model with GDELT performs much better for Mexico (with an AUC of 0.78 compared to 0.60 obtained with ICEWS). Also for Brazil, Paraguay and Venezuela, the models with GDELT outperform ICEWS. When we look at the ROC curves in Figure 19, we observe that combining the two databases improves the performance (higher median) compared to that obtained from each source separately for all countries except for Argentina.

Note that for some countries, the performance obtained by ICEWS or GDELT is as low as the model with Tor and currency (which has the lowest AUC compared to all other combinations for all countries). As seen in Figure 18, ICEWS database results in the worst performance for Mexico, and the model with GDELT results in the lowest AUC values for Colombia.

8.3 Tor and Currency

These two data sources are much less informative compared to others, resulting in the lowest AUC values, especially for countries such as Argentina and Paraguay. However, for some countries performance is improved by adding these data sources to political event databases (see Figure 18). For Brazil and Venezuela, the model with Tor and currency result in higher AUC values compared to other countries because Tor and currency are important features selected by Lasso in the all-data model, as shown in Figure 10 (and discussed in Section 7).

To sum up, the ablation study shows that social media has more predictive power than other data sources, and enhances the prediction performance. However, it also increases the variation; the performance metrics have wider ranges. Political event databases are not as informative as social media, but the ROC curves generated based on ICEWS and GDELT do not vary much compared to those based on social media.

9 Conclusions and Future Work

We present and implement a new approach for integrating multiple data sources to predict civil unrest in six Latin American countries. We also evaluate the predictive power of disparate datasets and methods, and provide interpretable insights into unrest events.

We observe that social media data such as Twitter, blogs, and news play an important role in predicting insurgencies. Political event databases, despite being more focused on political events, and hence expected to have more predictive ability, do not seem to perform as well as social media. We should acknowledge that there are differences in the generation and collection of these data sources, and our frequency-based prediction approach does not take into account these differences (increase of a keyword volume on Twitter is different than an

increase in the number of a particular event in the ICEWS database, which is an automated data source). Further research is needed to develop methods that take into account differences in data sources.

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

Acknowledgments

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, NSF NETSE Grant CNS-1011769 and NIH 1R01GM109718. 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. Arias M, Arratia A, Xuriguera R. Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2013; 5(1):8:1–8:24.
2. Asur S, Huberman BA. Predicting the future with social media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. 2010; 1:492–499.
3. Bellemare MF. Rising food prices, food price volatility, and social unrest. *American Journal of Agricultural Economics*. 2015; 97(2):1–21.
4. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *Journal of Computational Science*. 2011; 2(1):1–8.
5. Chakraborty P, Khadivi P, Lewis B, Mahendiran A, Chen J, Butler P, Nsoesie EO, Mekaru SR, Brownstein JS, Marathe M, et al. Forecasting a moving target: Ensemble models for ILI case count predictions. *SIAM Data Mining*. 2014:262–270.
6. Chen F, Neill DB. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs; *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2014. 1166–1175.
7. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the first Workshop on Social Media Analytics*. 2010:115–122.
8. El-Katiri L, Fattouh B, Mallinson R. The Arab uprisings and MENA political instability: Implications for oil & gas markets. *Oxford Institute for Energy Studies*; 2014.
9. Gerner DJ, Schrodt PA, Francisco RA, Weddle JL. Machine coding of event data using regional and international sources. *International Studies Quarterly*. 1994:91–119.
10. Gerner DJ, Schrodt PA, Yilmaz O, Abu-Jabr R. Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. *43rd Annual Convention of the International Studies Association*. 2002:24–27.
11. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. *Numerische Mathematik*. 1970; 14(5):403–420.
12. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y. The dynamics of protest recruitment through an online network. *Scientific Reports*. 2011; 1(197)
13. Kallus N. Predicting crowd behavior with big public data; *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*; 2014. 625–630.
14. Keneshloo Y, Cadena J, Korkmaz G, Ramakrishnan N. Detecting and forecasting domestic political crises: A graph-based approach; *Proceedings of the 2014 ACM Conference on Web Science*; 2014. 192–196.

15. Korkmaz G, Cadena J, Kuhlman CJ, Marathe A, Vullikanti A, Ramakrishnan N. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM; 2015. Combining heterogeneous data sources for civil unrest forecasting; 258–265.
16. Lampos V, De Bie T, Cristianini N. Machine Learning and Knowledge Discovery in Databases. Springer; 2010. Flu detector-tracking epidemics on Twitter; 599–602.
17. Leetaru K, Schrodt PA. International Studies Association (ISA) Annual Convention. Vol. 2. CiteSeer; 2013. GDELT: Global data on events, location, and tone, 1979–2012.
18. Lynch J. The Spanish-American Revolutions, 1808–1826. Weidenfeld & Nicolson; 1973.
19. McFadden D. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*. 1973;105–142.
20. McFadden D. Tech. rep., Cowles Foundation for Research in Economics. Yale University; 1977. Quantitative methods for analyzing travel behaviour of individuals: Some recent developments.
21. Muthiah S, Huang B, Arredondo J, Mares D, Getoor L, Katz G, Ramakrishnan N. Planned protest modeling in news and social media; Proceedings of the Twenty-Seventh Annual Conference on Innovative Applications of Artificial Intelligence (IAAI); 2015. 3920–3927.
22. Piven FF, Cloward RA. Poor People's Movements. Pantheon; 1977.
23. Ramakrishnan N, Butler P, Muthiah S, Self N, Khandpur R, Saraf P, Wang W, Cadena J, Vullikanti A, Korkmaz G, Kuhlman C, Marathe A, Zhao L, Hua T, Chen F, Lu CT, Huang B, Srinivasan A, Trinh K, Getoor L, Katz G, Doyle A, Ackermann C, Zavorin I, Ford J, Summers K, Fayed Y, Arredondo J, Gupta D, Mares D. 'beating the news' with EMBERS: Forecasting civil unrest using open source indicators; Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2014. 1799–1808.
24. Sakaki T, Okazaki M, Matsuo Y. Proceedings of the 19th International Conference on World Wide Web. ACM; 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors; 851–860.
25. Shi L, Agarwal N, Agrawal A, Garg R, Spoelstra J. Predicting US primary elections with Twitter. 2012. URL: <http://snap.stanford.edu/social2012/papers/shi>
26. Starbird K, Palen L. (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising; Proceedings of the 2012 ACM Conference on Computer Supported Cooperative Work; 2012. 7–16.
27. Stoll RJ, Subramanian D. Hubs, authorities, and networks: Predicting conflict using events data; The International Studies Association Conference; 2006.
28. Székely GJ, Rizzo ML, Bakirov NK, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*. 2007; 35(6):2769–2794.
29. Tang J, Wang X, Liu H. Modeling and Mining Ubiquitous Social Media. Springer; 2012. Integrating social media data for community detection; 1–20.
30. Theocharis Y. The wealth of (occupation) networks? Communication patterns and information distribution in a Twitter protest network. *Journal of Information Technology & Politics*. 2013; 10(1):35–56.
31. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996;267–288.
32. Tumasjan A, Sprenger TO, Sandner PG, Welpel IM. Predicting elections with Twitter: What 140 characters reveal about political sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM). 2010; 10:178–185.
33. Ward MD, Metternich NW, Carrington C, Dorff C, Gallop M, Hollenbach FM, Schultz A, Weschle S. Geographical models of crises: Evidence from ICEWS. *Advances in Design for Cross-Cultural Activities*. 2012;429.
34. Wulf V, Aal K, Abu Kteish I, Atam M, Schubert K, Rohde M, Yerosus GP, Randall D. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2013. Fighting against the wall: Social media use by political activists in a Palestinian village; 1979–1988.
35. Wulf V, Misaki K, Atam M, Randall D, Rohde M. 'On the ground' in Sidi Bouzid: Investigating social media use during the Tunisian revolution; Proceedings of the 2013 ACM Conference on Computer Supported Cooperative Work; 2013. 1409–1418.

36. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68(1):49–67.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1. Several data sources are used in a Lasso regression to forecast civil unrest events. The data sources include Twitter, news, blogs, Tor (The Onion Router), political event databases (ICEWS and GDELT), and exchange rate.

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. 2.
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.

Fig. 3. Word cloud of the civil unrest dictionary used for filtering Twitter, news, and blogs. It includes *protest-related words* (“gobierno” (government), “estudiantes” (students)), *key phrases* (“salir a la calle” (take the streets)), country-specific *key players* (“Henrique Capriles”).

Fig. 4.

Normalized time series of keyword counts in Twitter messages (left) 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.

Fig. 5.

Box plots of the distributions of F1-scores for Baseline (red), Lasso (green) and Hybrid (blue) models for 300-day training periods and testing periods of 10, 20 and 30 days, respectively.

Fig. 6.

The ROC curves corresponding to the median AUC (given in parentheses) over 50 executions of the GLM model. The AUC values correspond to the median AUC's given in Table 10 above. The distribution summary of the AUC's for all executions (min, mean, and max) is also given in Table 10.

Fig. 7.

Estimated probabilities, $\hat{Y} \equiv \hat{P}_{t+1}$, compared to the response variables, $Y \equiv y_{t+1}$, of the GLM model with median AUC. GLM model estimates higher probabilities when the response variable is 1. There is more variance when $Y=0$, which means that the model is better at correctly identifying when there is going to be an event but somewhat prone to emitting false positives.

Fig. 8.

Box plots showing the entire values of AUCs over 50 executions of (i) Lasso, and (ii) GLM model (logistic regression with Lasso-selected variables).

Fig. 9.
Distribution of GSR events in Brazil, Mexico and Venezuela (November 1st, 2012 – August 31st, 2014).

Fig. 10.

Word clouds of the features selected by Lasso for Brazil, Mexico, and Venezuela. The keywords that are not indicated as “news” or “blogs” are features from Twitter. The size of the features indicates the frequency of selection by Lasso over the 50 executions. The maximum frequency is 50 (in dark blue), and the minimum frequency is chosen as 10.

Fig. 11.

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 date format is YY-MM-DD. The upward trend corresponds to the start of nationwide protests in Brazil in June 2013 (Brazilian Spring).

Fig. 12.
The frequency of the keyword “comerciante”(left) and “racismo” (right) in news sources.
The peaks correspond to the dates of GSR events in Figure 13.

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. 13.

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 Lasso from social media (see Table 8 and Figure 10). The first two entries correspond to Brazilian Spring, June 2013 discussed in Section 7.1. The last event is discussed in Section 7.2.

Fig. 14.
Venezuelan protesters sign: "Why do I protest? Insecurity, scarcity, injustices, repression, deceit, for my future."

Fig. 15.

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 date format is YY-MM-DD. The increased use of the word corresponds to the dates of student-led protests in Venezuela in February 2014.

Fig. 16.

The time series of the volume of the word “protesta” (protest) in Twitter, news and blogs compared to the GSR (in dark green). The keyword “protesta” is an important indicator of civil unrest in Mexico. However, the usage of this keyword in blogs (right) provides a better signal than Twitter (left) or news (center). The volume of the keyword in blogs captures the rise in activity early in 2014 and the “dip” in July of the same year.

Fig. 17.

The time series of the total volume of keywords in Twitter (top-left) and blogs (top-right) in Mexico, and the number of GSR events (in green). The reduction in the GSR events in July 2014 is followed by the same trend in the blog activity (but not on Twitter). Combining the data sources provide additional information to avoid false positives, reflected as a decline in the probabilities generated by our model (the blue series in the bottom chart).

Fig. 18.

The values of AUC for the ablation study; the values for 50 executions of Lasso under different combinations of data sources are shown.

Fig. 19.

The ROC curves for the ablation study; the curves corresponding to the median AUC (value given in parenthesis) over 50 executions of Lasso are shown. The performance varies with different combinations of data sources.

Fig. 20.

The ROC curves for the ablation study; the curves for 50 executions of Lasso under different combinations of data sources are shown. The median AUC is given in parenthesis. The performance varies with different combinations of data sources.

Table 1

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

Table 2

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

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Summary of CAMEO codes used in ICEWS and GDELT.

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

Table 4

Number of features per data source.

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

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Number of observations and days with GSR events for each country.

Country	# Observations	# Events	Country	# Observations	# Events
Argentina	545	268	Brazil	554	448
Colombia	507	227	Mexico	555	480
Paraguay	441	253	Venezuela	529	361

Table 6

Performance metrics of Lasso for 300-day training periods and a 5-day moving window. The metrics are averaged over the test periods.

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.

Table 7

Performance metrics of keyword-based Group-Lasso for 300-day training periods.

Country	Test Days	Precision	Recall	F1-score
Brazil	$T = 10$	0.82	0.79	0.77
	$T = 20$	0.81	0.86	0.79
	$T = 30$	0.81	0.81	0.76
Mexico	$T = 10$	0.78	1.00	0.88
	$T = 20$	0.92	0.83	0.83
	$T = 30$	0.88	0.91	0.85
Venezuela	$T = 10$	0.81	0.77	0.72
	$T = 20$	0.82	0.69	0.65
	$T = 30$	0.85	0.51	0.50

Table 8

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

Table 9

Summary of goodness-of-fit for GLM model (logistic regression with Lasso-selected variables). The median values over 50 executions are reported.

Country	Null Dev.	Resid. Dev.	Null DF	Resid. DF	$R^2_{McFadden}$	AUC
Argentina	755.38	485.73	544.00	507.00	0.36	0.87
Brazil	540.87	246.41	553.00	514.00	0.54	0.93
Colombia	697.30	632.14	506.00	502.00	0.09	0.69
Mexico	439.60	257.16	554.00	533.00	0.41	0.88
Paraguay	601.74	587.58	440.00	439.00	0.02	0.55
Venezuela	661.28	380.44	528.00	511.00	0.24	0.80

Number of Lasso-selected features and the performance of GLM model. The distribution summary (min, median, mean, and max) over 50 executions is reported.

Table 10

Country	Number of feat.	AUC	Country	Number of feat.	AUC	R^2_{McF}	AUC
Argentina (n=545)	Min	0.22	Mexico (n=555)	Min	0.34	0.85	
	Med.	0.36		Med.	0.41	0.88	
	Mean	0.34		Mean	0.45	0.89	
	Max	0.43		Max	0.69	0.96	
Brazil (n=554)	Min	0.40	Paraguay (n=441)	Min	0.02	0.55	
	Med.	0.54		Med.	0.02	0.57	
	Mean	0.56		Mean	0.03	0.58	
	Max	0.90		Max	0.11	0.71	
Colombia (n=507)	Min	0.06	Venezuela (n=529)	Min	0.16	0.74	
	Med.	0.09		Med.	0.24	0.80	
	Mean	0.34		Mean	0.27	0.82	
	Max	1.00		Max	0.47	0.92	

Table 11

The case studies for Brazil, Mexico and Venezuela.

Country	Event	Date	Lasso-selected features
Brazil	Brazilian Spring	June, 2013	Currency, “comerciante” (news)
	Racism	January, 2014	“racismo” (news)
Mexico	General	Jan. – June, 2014	“protesta” (blogs, news and Twitter)
	General	July, 2014	Twitter and blogs
Venezuela	Student protests	February, 2014	Tor, “represión” (Twitter)