EMBERS at 4 years: Experiences operating an Open Source Indicators Forecasting System

Sathappan Muthiah[*], Patrick Butler[*], Rupinder Paul Khandpur[*], Parang Saraf[*], Nathan Self[*], Alla Rozovskaya[*], Liang Zhao[*], Jose Cadena[†], Chang-Tien Lu[*], Anil Vullikanti[†], Achla Marathe[†], Kristen Summers[‡], Graham Katz[§], Andy Doyle[§], Jaime Arredondo[¶], Dipak K. Gupta[‖], David Mares[¶], Naren Ramakrishnan[*]

April 4, 2016

Abstract

EMBERS is an anticipatory intelligence system forecasting population-level events in multiple countries of Latin America. A deployed system from 2012, EMBERS has been generating alerts 24x7 by ingesting a broad range of data sources including news, blogs, tweets, machine coded events, currency rates, and food prices. In this paper, we describe our experiences operating EMBERS continuously for nearly 4 years, with specific attention to the discoveries it has enabled, correct as well as missed forecasts, and lessons learnt from participating in a forecasting tournament including our perspectives on the limits of forecasting and ethical considerations.

1 Introduction

Modern communication forms such as social media and microblogs are not only rapidly advancing our understanding of the world but also improving the methods by which we can comprehend, and even forecast, the progression of events. Tracking population-level activities via 'massive passive' data has been shown to quite accurately shed light into large-scale societal movements.

Two years ago, in KDD 2014, we described EMBERS [15], a deployed anticipatory intelligence system [5] that forecasts significant societal events (e.g., civil unrest events such as protests, strikes, and 'occupy' events) using a large set of open source indicators such as news, blogs, tweets, food prices, currency rates, and other public data. The EMBERS system has been running continuously 24x7 for nearly 4 years at this point and our goal in this paper is to present the discoveries it has enabled, both correct as well as missed forecasts, and lessons learned from participating in a forecasting tournament including our perspectives on the limits of forecasting and ethical considerations. In

[*] Discovery Analytics Center, Virginia Tech, Arlington,VA 22203
[†] Biocomplexity Institute, Virginia Tech, Blacksburg, VA 24061
[¶] University of California at San Diego, San Diego, CA 92093
[‖] San Diego State University, San Diego, CA 92182
[§] CACI Inc., Lanham, MD 20706
[‡] IBM Watson Group, Chantilly, VA 20151

particular, we shed insight into the value proposition to an analyst and how EMBERS forecasts are communicated to its end-users.

The development of EMBERS is supported by the Intelligence Advanced Research Projects Activity (IARPA) Open Source Indicators (OSI) program. EMBERS forecasts are scored against the Gold Standard Report (GSR), a monthly catalog of events as reported in newspapers of record in 10 Latin American countries - Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, Venezuela. The GSR is compiled by MITRE corporation using human analysts. EMBERS currently focuses on multiple regions of the world but for the purpose of this paper we focus primarily on Latin America, specifically the countries of Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. Similarly, EMBERS generates forecasts for multiple event classes—influenza like illnesses [3], rare diseases [16], elections [12], domestic political crises [8], and civil unrest—but in this paper we focus primarily on civil unrest as this was the most challenging event class with hundreds of events every month across the countries studied here.

Our key contributions can be summarized as follows:

1. Unlike retrospective studies of predictability, EMBERS forecasts are communicated in real-time before the event to MITRE/IARPA and scored independently of the authors. We present multiple quantitative indicators of EMBERS performance as well as insights into how we made EMBERS forecasts most valuable to creazione di una We report two primary ways in which analysts utilize EMBERS and the use of *automated narratives* to help make EMBERS forecasts as useful as possible.

2. In an attempt to demystify the state-of-the-art in forecasting and to create an open dialogue in the community, we report both successful forecasts of EMBERS as well as events missed by EMBERS. The events not forecast by EMBERS lead us to considerations of both the limitations of the underlying technology as well as the inherent limits to forecasting large-scale events.

3. While social media is often touted as the key to event forecasting systems such as EMBERS, we present the results of an ablation study to outline the performance degradation that ensues if data sources such as Twitter and Facebook were to be removed from the forecasting pipeline.

4. We consider the separation of civil unrest events into events that happen with a degree of regularity versus rare or significant events, and evaluate the performance of EMBERS in forecasting such surprising events.

5. We describe our current best understanding of the limitations to forecasting civil unrest events using technologies like EMBERS and also consider the ethical considerations of the EMBERS technologies.

2 Background

We begin by providing a brief review of forecasting systems, followed by a quick preview of EMBERS, its system architecture, machine learning models, and measures for

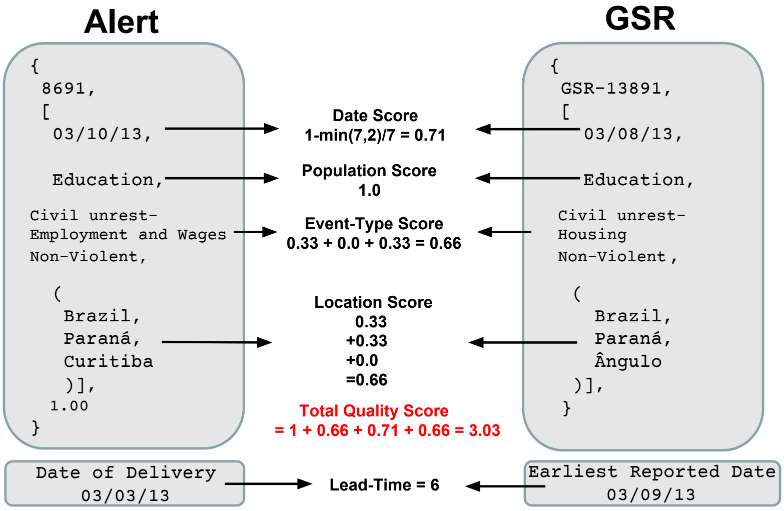


Figure 1: An example depicting how an alert is scored with respect to the ground truth.

evaluating its performance. For more details, please see [15].

Forecasting societal events such as civil unrest has a long tradition in the intelligence analysis and political science community. We distinguish between forecasting systems versus event coding systems (systems that provide structured representations of ongoing events reported in newspapers), and focus on the former. Early forecasting systems such as ICEWS [14] provided very broad coverage in countries but were limited by their spatio-temporal resolution (e.g., typically country- and month- level forecasting for specific events of interest [19]). The ICEWS events of interest are domestic political crises, international crises, ethnic/religious violence, insurgencies, and rebellion. A similar project in scope is PITF (Political Instability Task Force) [6] funded by the CIA. To the best of our knowledge, only EMBERS provides the most specific spatial resolution (city-level) and temporal resolution (daily-level) capability in forecasting.

The software architecture of EMBERS (Early Model Based Event Recognition using Surrogates) is designed as a loosely coupled, share-nothing, highly distributed pipeline of processes connected via ZeroMQ. In this manner, the system is both highly scalable and fault tolerant. The EMBERS pipeline can loosely be broken up into four stages: ingestion, enrichment, modeling, and selection. In the first stage, ingestion, data is collected from a variety of sources and streamed into the following stages in real-time. The enrichment stage takes the raw data from the ingestion stage and processes it in various ways including natural language processing, geocoding, and relative time phrase normalization. After enrichment, the modeling stage feeds the enriched data into the various models that make up EMBERS. Unlike other systems which use single monolithic models to make predictions, EMBERS combines the results of several different models to arrive at the most accurate forecasts. Finally, in the selection stage the separate alerts from each model are de-duplicated, fused, and selected and finally emitted as a full forecast for a real world event.

The structure of a civil unrest forecast is shown in Figure 1 (left). A forecast constitutes four fields, corresponding to the when, where, who, and why of the protest. These fields are respectively denoted as the date, location, population, and event type. Location is recorded at the city level. Population and event type are fields chosen from a categorical set of possibilities. The figure further shows how an alert with all these fields are scored against a GSR event. In the basic scoring methodology shown in Figure 1 each of the four fields are weighted uniformly and a total quality score out of 4 is obtained. Apart from this each alert also has a lead-time associated with it calculated as shown in Figure 2.

Rather than design one model to integrate all possible data sources, EMBERS adopted a multi-model approach to forecasting. Each model utilized a specific (possibly overlapping) set of data sources and is tuned for high precision, so that the union of these models can be tuned for high recall. A fusion/suppression engine [7] allows a tunable strategy to issue more or fewer alerts depending on whether the analyst's objective is to obtain a higher precision or recall. The underlying models used in EMBERS are: (i) *planned protest model* [13], (ii) *dynamic query expansion* [20], (iii) *volume-based model* [9], (iv) *cascade regression* [2], and (v) a baseline model. The planned protest model, for news and social media (Twitter, Facebook), identifies explicit signs of organization and calls for protest, resolves relative mentions of time (e.g., 'next Saturday') and space (e.g., 'the square') to issue forecasts. The dynamic query expansion (DQE) model uses Twitter as a data source and learns time- and country-specific expansions of a seed set of keywords to identify specific situational circumstances for civil unrest. For instance, in Venezuela (an economy where the government exercises stringent price controls), there were a series of protests in 2014 stemming from the shortage of toilet paper, a novel circumstance that was uncovered by DQE. The volume-based model uses a range of data sources, spanning social, economic and political indicators. It uses classical statistical models (LASSO and hybrid regression models) to forecast civil unrest events using features from social media (Twitter and blogs), news sources, political event databases (ICEWS and GDELT [10]), Tor [4] statistics, food prices, and currency exchange rates. It aims to provide a multi-source perspective into forecasting by leveraging the selective superiorities of different data sources. The cascade regression model aims to model activity related to organization and mobilization in Twitter [2]. Finally, the baseline model uses a maximum likelihood estimation over the GSR to issue history-based forecasts.

The EMBERS project is unique not just in its algorithmic underpinnings but also in the use of new measures for evaluation, specifically aimed at determining forecasting performance. As shown in Figure 2, one of the primary measures of EMBERS performance is lead time, the number of days by which a forecast 'beats the news', i.e., the date of reporting of the event. Lead time should not be confused with date quality, i.e., the difference between the predicted date and the actual date of the event. The date quality is one of the components to the quality score, the other components being the location score, event type score, and population score. Figure 1 shows how these other components are scored between an EMBERS forecast and a GSR record. Given a set of alerts and a set of GSR events for a given month, the lead time is used as a constraint to

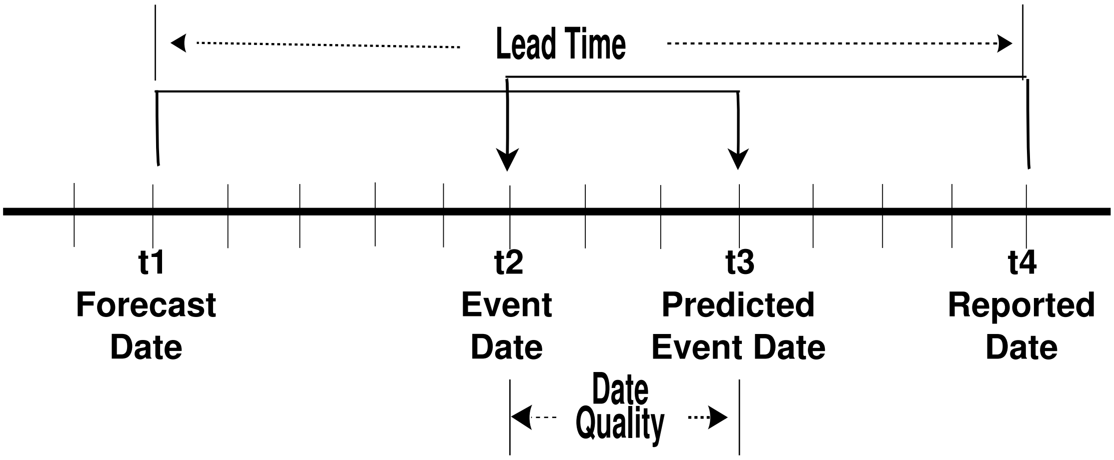


Figure 2: Alert sent at time $t1$ predicting an event at time $t3$ can be matched to a GSR event that happened at time $t2$ and reported at time $t4$ if $t1 < t4$.

define legal (alert, event) pairs so that we can construct a bipartite matching to optimize the best quality score. From this bipartite matching, measures of precision and recall can be derived, i.e., by assessing the number of (un)matched events or alerts. Finally, a confidence score is used to assess the quality of probabilities imputed by EMBERS to its forecasts, and measured in terms of the Brier score. For more details, please see [15].

We now turn to a discussion of specific discoveries enabled by EMBERS, into civil unrest in Latin America, and into the complexity of the forecasting enterprise as a whole.

3 Performance Analysis

First, we begin with a performance analysis of EMBERS, from both a quantitative point of view with respect to the GSR and with respect to end-user (analyst) goals.

3.1 Quantitative Metrics

Figure 3 depicts both the targets set by the IARPA OSI program as well as the actual measures achieved by the EMBERS system. As shown here, the easiest target to achieve in EMBERS was, surprisingly, the lead time objective. This was feasible due to EMBERS's focus on modeling both planned and spontaneous events. Planned events are sometimes organized with as many as several weeks of lead time and thus identifying indicators of organization was instrumental in achieving lead time objectives. The confidence (mean probability) scores were also achieved by EMBERS and involved careful calibration of probabilities by taking into account estimates of model propensities and data source reliabilities. The measure that was most difficult to achieve was the quality score as it involved a four component additive score and thus tangible improvements in score required more than incremental improvements in forecasting specific components.

Targets

	Month 12	Month 24	Month 36
Mean Lead-Time	1 day	3 days	7 days
Mean Probability Score	0.60	0.70	0.85
Mean Quality Score	3.0	3.25	3.5
Recall	0.50	0.65	0.80
Precision	0.50	0. 65	0.80

Actual

Metric	Month 12	Month 24	Month 36
Mean Lead-Time	3.89 days	7.54 days	9.76 days
Mean Probability Score	0.72	0.89	0.88
Mean Quality Score	2.57	3.1	3.4
Recall	0.80	0.65	0.79
Precision	0.59	0.94	0.87

Figure 3: IARPA OSI targets and results achieved by EMBERS

Finally, recall and precision involve a natural underlying trade-off and the deployment of our fusion/suppression engine provided the ability to balance this trade-off to meet IARPA OSI's objectives.

Apart from comparing mean scores another interesting metric is to see how many perfect matches (4.0 quality score) are obtained by an algorithm.Figure 4 shows the number of alerts issued by EMBERS that matched perfectly to an event in the future on a monthly basis for 2013. The figure clearly shows that EMBERS makes almost double the number of fully accurate forecasts as compared to the baserate model.

3.2 Analyst Evaluation

In addition to the quantitative measures above, our experience interacting with analysts (across multiple branches of government) demonstrated an interesting dichotomy as to how analysts use EMBERS alerts. Some analysts preferred to use EMBERS in an 'analytic triage' scenario wherein they could tune EMBERS for high recall so that they would apply their traditional measures of filtering and analysis to hone in on forecasts of interest. Other analysts instead viewed EMBERS as a data source and preferred to use it in a high precision mode, e.g., wherein they were focused on a specific region of the world (e.g., Venezuela) and aimed to investigate a particular social science hypothesis (e.g., whether disruptions in global oil markets led to civil unrest).

To support these diverse classes of users, we implemented two mechanisms in the alert delivery stage. First, we implemented a mechanism wherein in addition to generating alerts, EMBERS also forecasted the expected quality score for each forecast (using machine learning methods trained on past GSR-alert matches). This expected quality

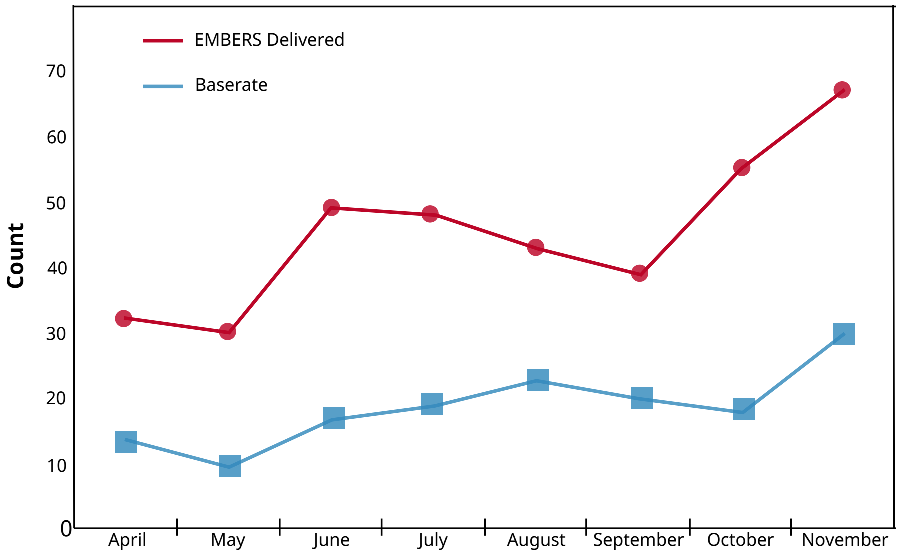


Figure 4: Comparison of number of perfect scores (4.0) obtained by EMBERS vs a baserate model each month in 2013.

score measure provided a way for analysts to use quality directly as a way to tune the system to receive greater or fewer alerts. Figure 5 shows the trade-off between final quality score and recall when alerts are suppressed based on expected quality. As expected we can see that the recall drops and quality increases as the cut-off threshold for expected quality is increased.

Second, we implemented an automated narrative generation capability (see Fig. 6) wherein EMBERS auto-generates a summary of the alert in English prose. As shown in Fig. 6, a narrative comprises many parts drawn from different sources of information. One source is the named entities and the system uses 'Wikification' to identify definitions and descriptions of these named entities on Wikipedia. A second source is historical (or real-time) statistics of warning output and warning performance and situating the alert in this context. The third source pertains to inferred reasons for the protest using knowledge graph identification techniques.

4 EMBERS Successes and Misses

Next, we detail some of the successful as well as not so successful forecasts made by EMBERS over the past few years in Latin America.

4.1 Successful Forecasts

Brazilian Spring (June 2013)

These protests were the largest and most significant protests in Brazil's recent history and caught worldwide attention. Millions of Brazilians took part in these demonstrations,

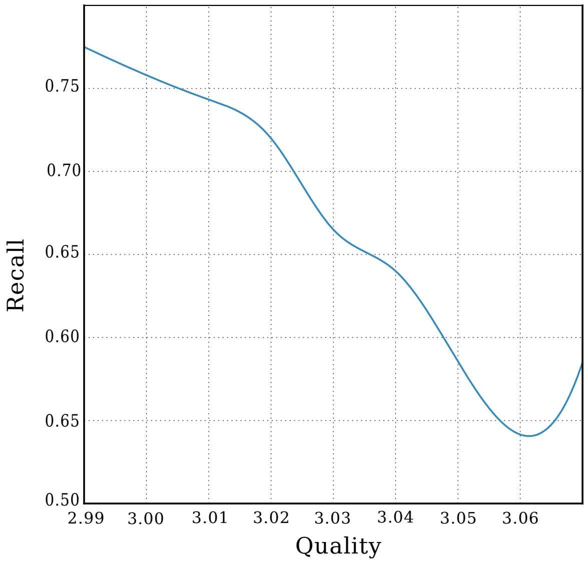


Figure 5: Recall vs quality tradeoff in EMBERS.

also known as the Brazilian Spring or the Vinegar Movement (inspired from the use of vinegar soaked cloth by demonstrators to protect themselves from police teargas). These protests were sparked by an increase in public transport fares from $R\$3$ to $R\$3.20$ by the government of President Dilma Rousseff.

As shown in Figure 7 EMBERS, while missing the initial uptick, forecast the increase in the order-of-magnitude of protest events during the Brazilian Spring and also captured the spatial spread in the events, in addition to forecasting that this event will span the broad Brazilian general population (as opposed to being confined to specific sectors).

Around 68% of EMBERS alerts during this period originated from the planned protest model. This is due to the fact that social networking platforms (Twitter and Facebook) as well as conventional news media played a key role in organization of these uprisings. Although initial protests were primarily due to the bus fare increases, they quickly morphed into more broader dissatisfaction to include wider issues such as government corruption, over-spending, and police brutality. The demonstrators also made calls for political reforms. In response, President Rousseff proposed a plebiscite on widespread political reforms in Brazil (but this was was later abandoned). Through its dynamic query expansion model, EMBERS was able to capture such discussions on Twitter (see Figure 8), and tracked their evolution as events unfolded through June.

The protests intensified in late June (see Figure 7), which were forecast correctly by EMBERS, and these events also coincided with FIFA 2013 Confederation Cup matches. We believe this was an important factor in helping the protests gain momentum, as the events were covered by international media. A majority of protests occurred in the cities that were hosting FIFA soccer matches. EMBERS issued most of its alerts for these host cities (see Figure 9), viz. Rio de Janeiro, São Paulo, Belo Horizonte, Salvador, and Porto Alegre, among others. For example, on 27th June during the Confederations Cup

EMBERS forecasts that there will be a violent protest on February, 18th 2014 in Caracas, the capital city of Venezuela. We predict that the protest will involve people working in the business sector. The protest will be related to discontent about economic policies. There were 5, 5, and 5 other similar warnings in last 2, 7 and 30 days, respectively.
The forecast date of the warning falls in week 7, which may have historical importance; this week is found to be statistically significant (pval=0.00461919415894, zscore=2.832, avg. count=57.25, mean=21.569 +/- 12.597)
Audit trail of the warning includes an article printed 2014-02-17.
Major players involved in the protest include Venezuelan opposition leader, students, President Nicolas Maduro, and Leopoldo Lopez.
Reasons: Protest against rising inflation and crime; Protestors want a political change; President Nicolas Maduro has accused US consular officials and right-wing.
Protests are characterized by: Venezuelan opposition leader spearheaded days of protest and calling for peaceful demonstration; Maduro accused official on 2014-12-16; Protests have seen several deadly street protests; Three people were killed on 2014-02-12; Demonstrations setting days of clashes; supporters to march to Interior Ministry on 2014-02-18.

Figure 6: An example narrative for a EMBERS alert message. Here, color red indicates named entities, green refers to descriptive protest related keywords. Items in blue are historical or real time statistics and those in magenta refer to inferred reasons of protest.

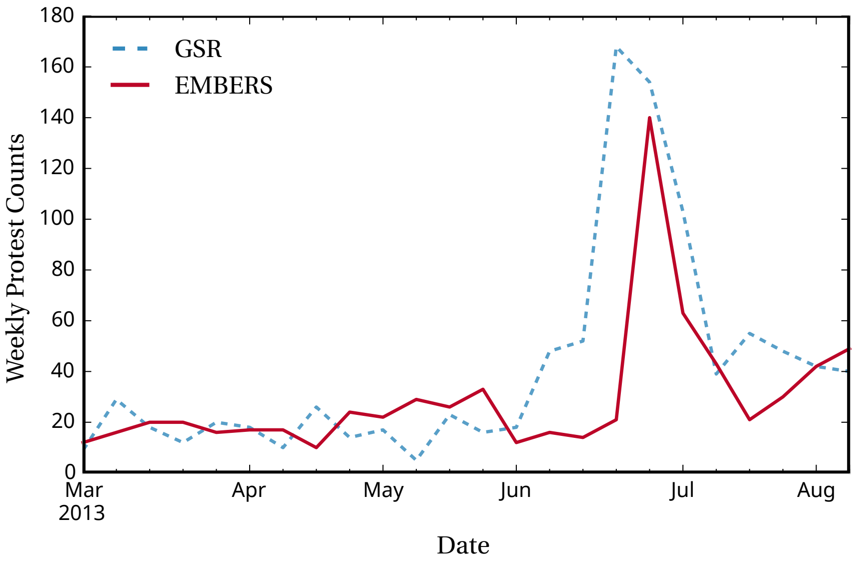


Figure 7: EMBERS performance during the Brazilian Spring (June 2013.)

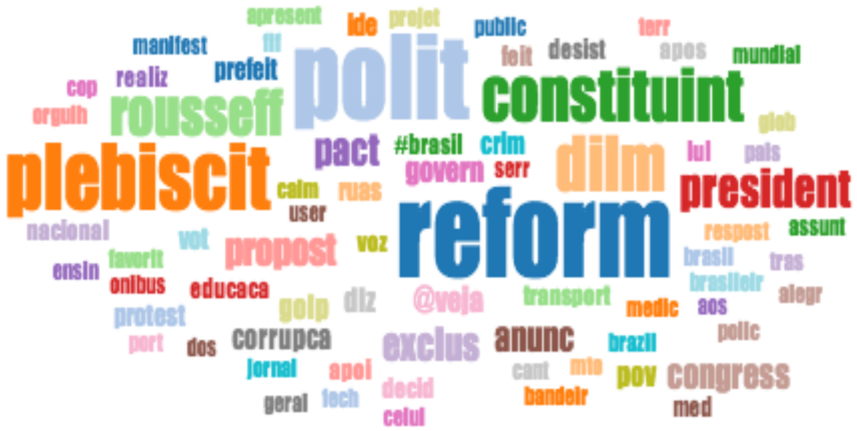


Figure 8: Word cloud representing tweets identified by EMBERS dynamic query expansion model.

semi-final in Fortaleza, around 5000 protesters clashed with the police near the Castelao stadium. In this case EMBERS had forecast an alert the day before. Later on the 30th of June, when the last games of the confederation cup took place in Rio de Janeiro and Salvador, they were plagued by mass protests as well; EMBERS predicted these events and submitted multiple alerts for Rio for the 28th and 29th June and one for Salvador for 29th June.

Venezuelan Protests (Feb-March 2014)

Early 2014 Venezuela started experiencing a situation of turmoil with a large portion of its population protesting due to insecurity, inflation and shortage of basic goods. This period saw one of the highest levels of civil disobedience in Venezuela with protests beginning in January with the murder of a former Miss Venezuela. However, the protests started gaining more importance and turned violent and more frequent with students joining the movement following an attempted rape of a student on campus in San Cristobal. EMBERS captured some of these first ‘calls to protest‘ at San Cristobal and its nearby surrounding areas and correctly forecast the population (Education) and that the protests would turn violent. A majority of the protesters were demanding that president Nicolas Maduro step down owing to the poor economic policies and widespread corruption. EMBERS was capable of capturing that the reason behind the protests were mainly against government policies with corruption being a major theme.The EMBERS models working on twitter were also clearly able to identify some of major leaders involved in the protest such as the major opposition leader Leopoldo Lopez. Though the events mainly started off from San Cristobal it spread widely throughout the country, EMBERS captured this spillover very well as shown in Figure 10. Figure 11 shows how EMBERS closely forecast the spike in the number of events during this period.

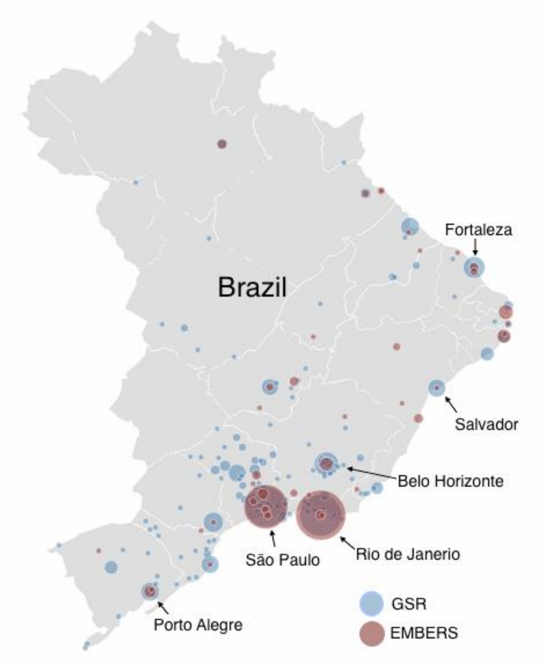


Figure 9: Geographic overlap of protest events (from the GSR) and EMBERS alerts for Brazil during June 2013.

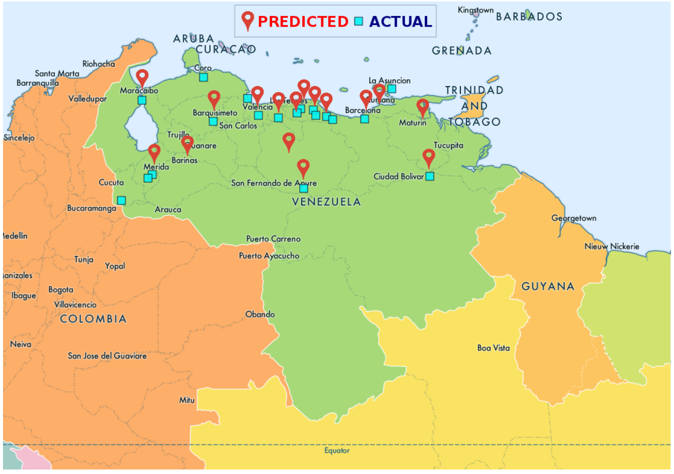


Figure 10: Geographical spread of protests (and forecasts) during Venezuelan student protests (Feb-Mar 2014).

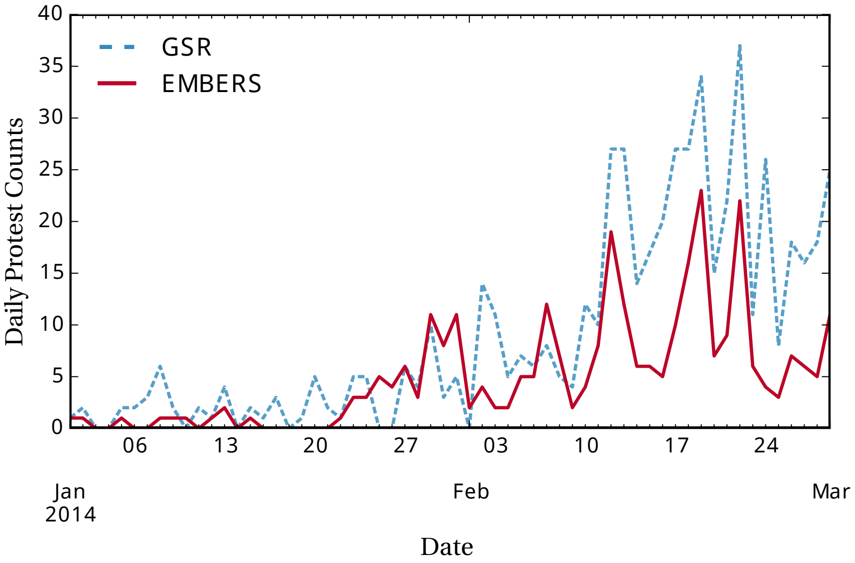


Figure 11: EMBERS performance during the Venezuelan student protests (Feb-Mar 2014).

Mexico Protests (Oct 2014)

Late September 2014, there were some peaceful protests by students from Ayotzinapa in Mexico against discriminatory hiring practices for teachers. During these protests, police opened fire on the students killing around three and about 43 went missing. This poor handling of the protest by the Mexican government caused widespread demonstrations throughout the country over the next few months in support of the families of the 43 missing students. A lot of these protests were violent in nature with demonstrators expressing extreme dissatisfaction against the government and president Pena Nieto. EMBERS, as shown in Figure 12 forecast an uptick in Mexico protests during early October 2014 with a lead time of about three days. It also generated a series of alert spikes coinciding with the first large-scale nationwide protests between October 5th and 8th. Figure 13 provides a timeline of GSR events and EMBERS alerts for Mexico during this period. This figure provides a detailed comparison of the continuous stream of alerts produced by EMBERS during this period with how the actual events unfurled in the real world.

Colombia Protests (Dec 2014 to March 2015

Colombia witnessed two different significant protests during this period, one during late December 2014 and the other during February 2015. Towards the end of 2014 the Colombian government were on the process of moving forward with peace negotiations to end 50 years of conflict with the Revolutionary Armed Forces of Colombia (FARC). With the FARC rebels having been associated with various acts of terror like extortion,

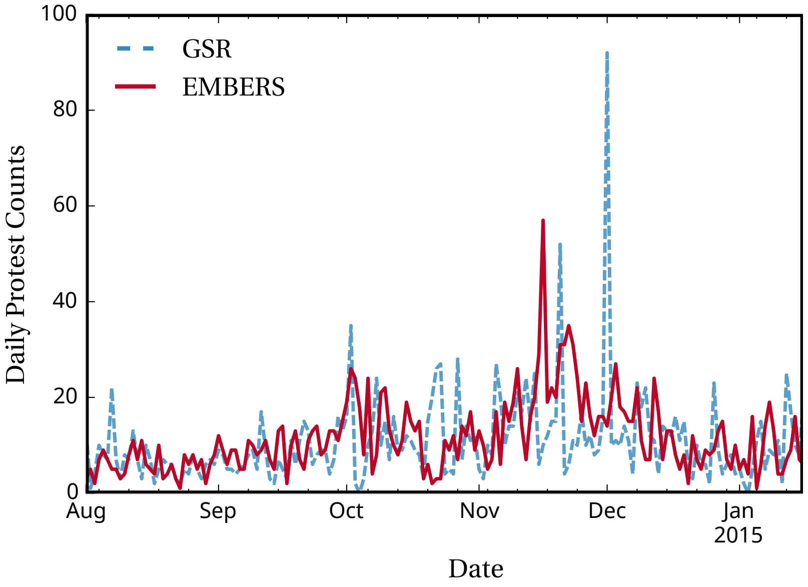


Figure 12: EMBERS performance during Mexico protests (Oct 2014).

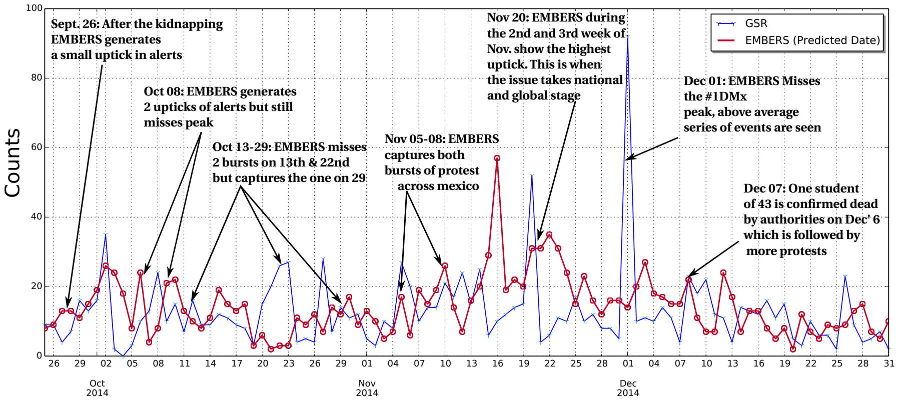


Figure 13: Timeline of Mexico protests, showing the correspondence between counts of GSR events and EMBERS alerts on a daily basis.

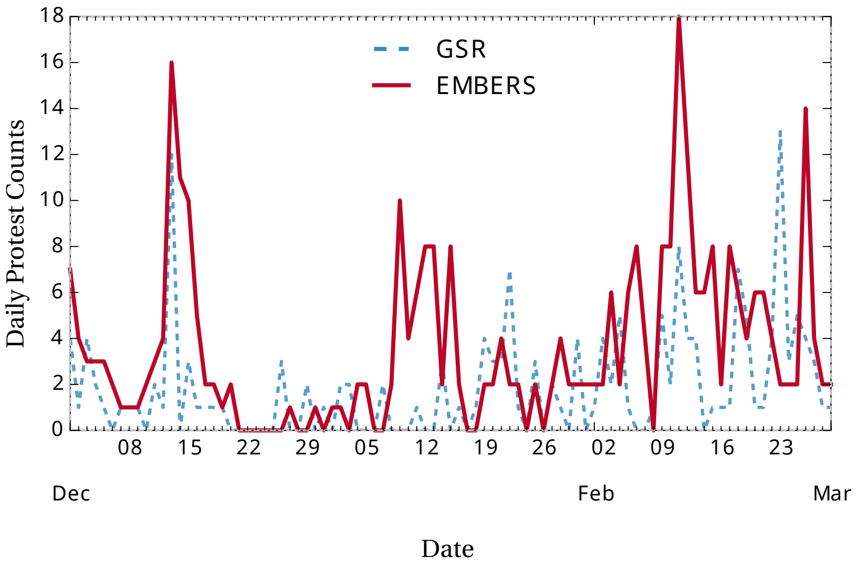


Figure 14: Embers Performance during the Colombia Protests (Dec.2014 to Mar.2015)

armed conflict, kidnapping, ransom, illegal mining etc., for a long period, the people of Colombia gathered in huge numbers to protest against possible amnesty for the FARC rebels.EMBERS successfully forecast the uptick in the number of events during the middle of December 2014 as indicated in Figure 14. The figure also shows the increase in protest counts during February 2015 though in this case EMBERS over-predicted the counts. The protests in February 2015 were mainly by led by truckers union demanding better freight rates, labor rights and against high fuel prices. The truckers extended for about a month and caused huge losses of about $300 million to the Colombian economy. EMBERS picked up on the truckers protests right from the beginning but was wrong in over estimating the numbers during February 11-12.

Paraguay Protests (February 2015)

The February 2015 protests in Paraguay were mainly carried out by peasants against the actions of the President Horacio Cartes.The protests were carried out after president Horacio's public revelation that he had opened two private swiss bank accounts. The protests also had a historical significance. It was also being carried out as a tribute to peasant leaders and activists who were murdered. The peasants also protested against the introduction of the new public-private partnership law.

EMBERS forecast the uptick in number of protest events in Paraguay during mid February 2015 as shown in Figure 15.

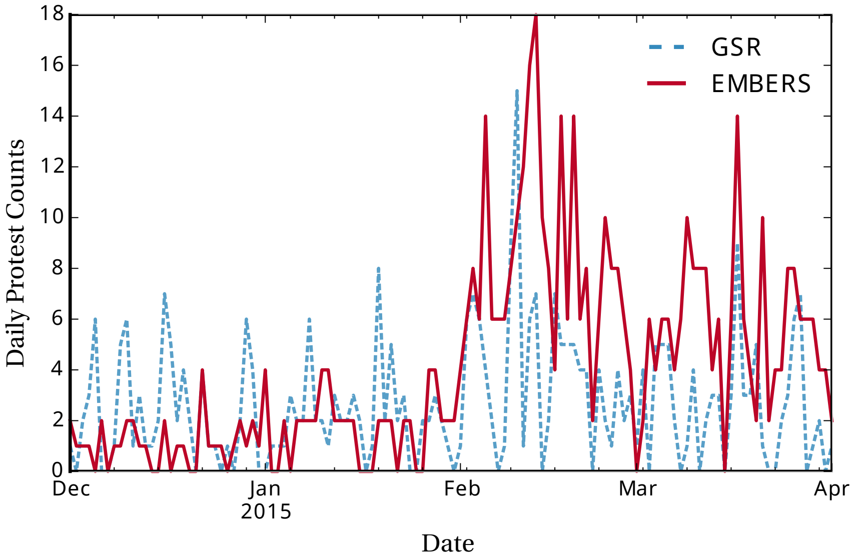


Figure 15: EMBERS performance during Paraguay Protests (Feb. 2015)

4.2 EMBERS Misses

Next, we outline specific large-scale events that EMBERS failed to forecast accurately, along with a discussion of underlying reasons.

4.2.1 Brazilian Protests (March 2015)

The beginning of 2015 saw a series of protests in Brazil demanding the removal of president Dilma Roussef amidst much furore against the increasing corruption in the country. The protest magnitude increased significantly due to the revelations that many politicians belonging to the ruling party accepted bribes from the state run energy company Petrobas. The protests saw huge participation from the general population, with protesters generally estimated to be around a million. EMBERS, as shown in Figure 16, picked up the onset of events but failed to capture the sudden rise in the number of events.

During this period there was a significant architectural change in the EMBERS processing pipeline. As mentioned in Section. 2 the EMBERS system enrichment pipeline consists of the following: natural language processing, geocoding and relative time phrase normalization (temporal tagging). During early 2015 EMBERS had moved to the Heideltime [17] temporal tagger from the previously used TIMEN [11] temporal tagger due to Heideltime's support for more languages and an active development cycle as opposed to TIMEN. Heideltime had no support for Portuguese (the primary language used in Brazil) and the EMBERS software development team had extended Heideltime to support Portuguese by translating the underlying resources for Spanish to Portuguese. As we learnt subsequently, the simple translation of rules from Spanish to Portuguese was

no portugese support after upgrade

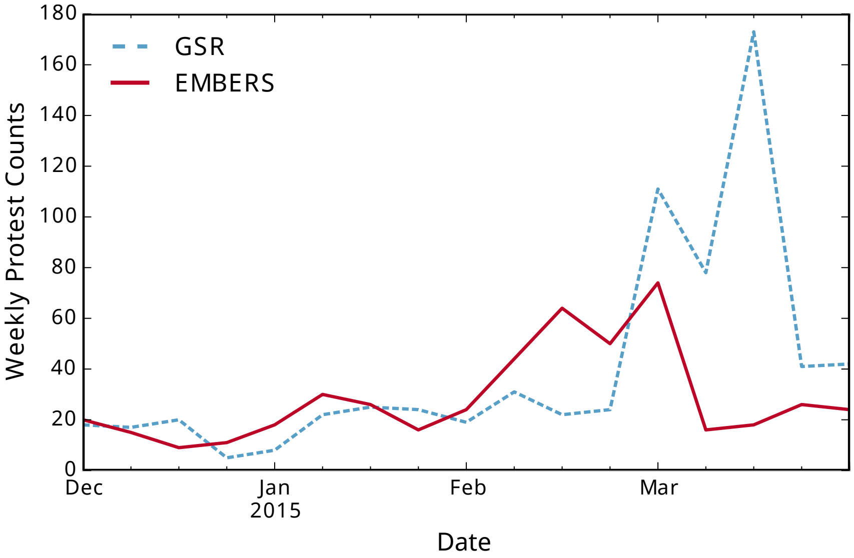


Figure 16: Brazil 2015 Protests

not sufficient and this affected the recall of one of the key models for Brazil, viz. the planned protest model. Since the planned protest model relies almost exclusively on the quality of information (specifically, date) extraction from text, its performance significantly deteriorated. This was subsequently corrected for the future.

4.2.2 *Mexico Protests (Dec 2014)*

The December 2014 were a continuation of the series of protests that began in October 2014 as described in Section. 4.1. People turned out in huge numbers in different cities of Mexico demanding President Pena Nieto's ouster owing to the manner in which the case of the 43 missing students were handled. The protests were largely peaceful except for a few cases where vehicles were torched and windows and office equipment were broken.

EMBERS missed the huge single day spike on December 1st. Though having predicted a nationwide event for December 1, EMBERS failed to capture the individual cities where the protest were carried out and also was unable to pick the magnitude. On manual analysis of the event, it was found that the date, December 1, was picked by the protesters due to its historical significance. This was the day when President Pena Nieto was sworn in as President in 2012 amidst much controversy and opposition from many specific constituent groups. The manual analysis also led to the understanding of how special dates were mentioned by twitterati **#1Dmx**. Dates mentioned like this were totally unrecognised by the EMBERS system and could have been one of the main reasons for EMBERS not being able to capture the peak on December 1st despite its historical significance.

Table 1: Comparison of performance metrics with and without social media sources. Social Media sources contributes a lot towards recall but loses out on lead-time.

Data Source	Quality-Score	Lead-time	Precision	Recall
Social Media Sources	-16.48%	-55%	+35%	-14%
Non-Social Media Sources	+8.42%	+30%	+79%	-33%

4.2.3 Brazilian Spring Onset

The EMBERS forecasts during the 2013 Brazilian Spring as shown in Fig. 7 was able to capture the peak but as can be seen the system was unable to capture the initial onset. See Section 7 for a detailed discussion of the limits of forecasting.

5 Ablation Testing

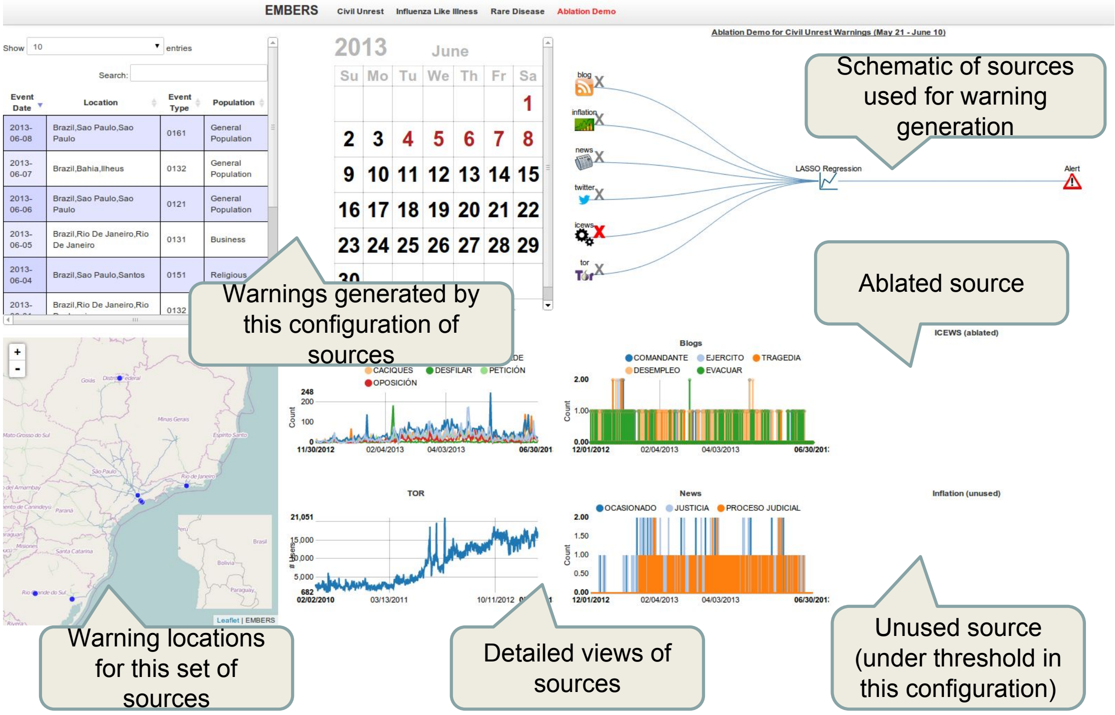
Different data sources provide different value to the forecasting enterprise. It is important we understand the value of a data source w.r.t. its forecasting potential. In this section we describe ablation testing in EMBERS where the value added to forecasting performance by traditional media sources like news and blogs are compared to that provided by the social media sources. Table 1 shows the percentage difference in the performance metrics namely quality score, lead-time, precision and recall when only one kind of source (traditional media or social media) is used with respect to the performance metrics when compared against the scenario when both sources are used. The table clearly shows that social media sources are mainly necessary in achieving high recall but not that useful in achieving high lead times for which the traditional media sources are required. This behavior is expected as social media is where daily chatter occurs whereas signs of organization and calls for protest often happen on traditional meadia. Mainly, Table 1 makes it evident that to build a successful forecasting system we need a good mix of both traditional and social media sources. Figure 17 shows a snapshot of the EMBERS ablation visualizer. The visualizer provides an analyst with the capability to selectively remove data sources and assess the differences in final alerts.

6 Forecasting Surprising Events

The GSR contains a mix of everyday, mundane, protests as well as surprising events such as the Brazilian Spring. We aimed to ascertain the relative ease of forecasting each class of events with respect to a simple baserate model.

The baserate model generates alerts using the rate of occurrence of events in the past three months.

To define surprising events, we employed a maximum entropy approach. For this purpose, each event is assumed to be describable in terms of three dimensions: country,



Event Date	Location	Event Type	Population
2013-06-08	Brazil,Sao Paulo,Sao Paulo	0161	General Population
2013-06-07	Brazil,Bahia,Ilheus	0132	General Population
2013-06-06	Brazil,Sao Paulo,Sao Paulo	0121	General Population
2013-06-05	Brazil,Rio De Janeiro,Rio De Janeiro	0131	Business
2013-06-04	Brazil,Sao Paulo,Santos	0151	Religious
2013-	Brazil,Rio De Janeiro,Rio	0132	

Figure 17: EMBERS visualization on Ablation testing

population and event type. The GSR can then be conceptualized as a cube. We infer a maximum entropy distribution conditioned on the marginals induced by the cube using iterative proportional fitting [1], as shown in Algorithm. 1.

Every month, events from the last three months of the GSR is used to populate the underlying cube of counts and the iterative proportional fitting procedure in Algorithm. 1 is used to estimate the maxent values for each cell. The resultant sample counts are then scaled to match the observed number of GSR events in the current month. The cell-wise difference between the inferred maxent distribution and the observed GSR is computed and all cells with significance greater than five standard deviations are classified as containing surprising events. In essence, this approach takes the GSR as input and creates a truncated GSR against which we can evaluate EMBERS (and the baserate model).

In Table 2 we present several inferred events in Latin America from our maximum entropy filter. For all these events, we compare the recall of both EMBERS and baserate model in Figure 18. It is clear that EMBERS is able to forecast these significant upticks consistently.

7 Uncertainties in Forecasting

While examining the events of civil unrest closely in the past few years, it was clear to our team that events carry two distinct types of uncertainties: **cause** and **timing**.

Table 2: Surprising events inferred by maxent analysis.

May ’13	Colombia	Nationwide protests against administrative policy changes regarding the Youth in Action program and social security pension
Jun ’13	Brazil	Brazilian Spring
Jul ’13	Brazil	Brazilian Spring
Sep ’13	Mexico	Nationwide protests against education and energy reforms
Oct ’13	Mexico	Nationwide protests against education and energy reforms
Oct ’13	Uruguay	Nationwide protests demanding increase in minimum wage
Feb ’14	Venezuela	Venezuelan Student Protests
Mar ’14	Venezuela	Venezuelan Student Protests
May ’14	Brazil	Nationwide demonstrations in response to the 2014 FIFA World Cup and other social issues
Jun ’14	Brazil	Nationwide demonstrations in response to the 2014 FIFA World Cup and other social issues
Aug ’14	Argentina	Nationwide protests against drops in wages, employment, and rising inflation
Sep ’14	Ecuador	Nationwide protests to demand changes in labor policies
Oct ’14	Mexico	Nationwide protests after the discovery of mass graves of kidnapped students

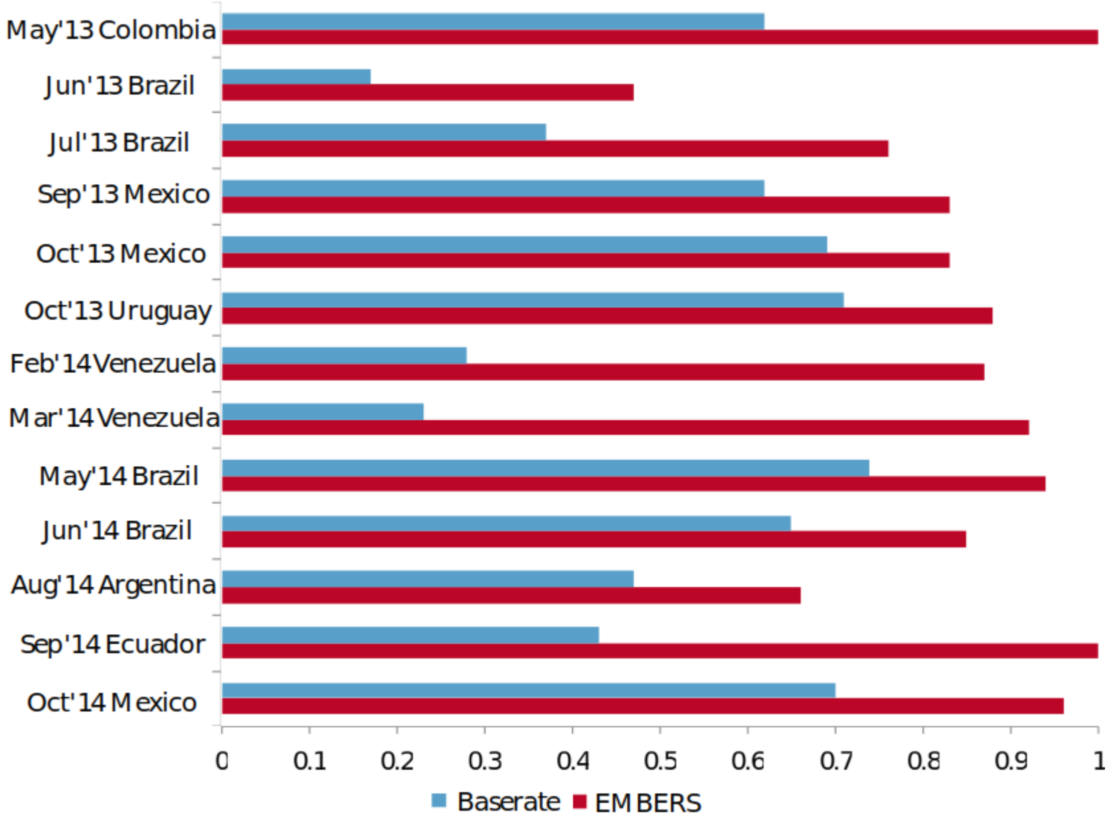
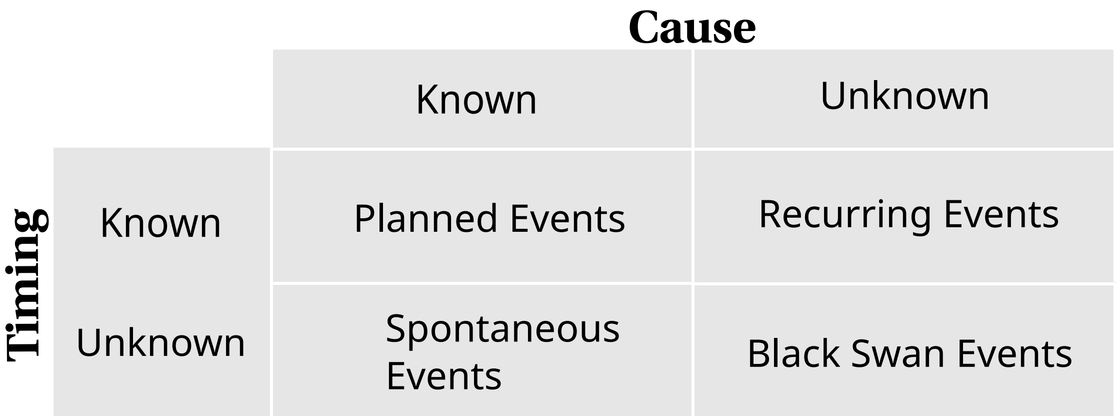


Figure 18: Performance of EMBERS vs a baserate model for surprising events.



		Cause	
		Known	Unknown
Timing	Known	Planned Events	Recurring Events
	Unknown	Spontaneous Events	Black Swan Events

Figure 19: Studying the limitations of event forecasting.

Algorithm 1 Surprise GSR calculation

```
procedure SURPRISE-GSR
    Input: $\mathcal{G} = \{\mathcal{G}_{-1}, \mathcal{G}_{-2}, \mathcal{G}_{-3}\}$
    Output: Maximum likelihood estimate for ¡event-type, population, country¿ tuple – $\mathcal{M}$
    Each event in the GSR-$\mathcal{G}$ is mapped to a three dimensional vector of ¡event-type, population, country¿. Each such vector is mapped to a cell in a 3-D cube and the value $x_{ijk}$ for a cell in the cube represents the frequency of the vector ¡i, j, k¿. $m_{ijk}$ represents the maximum likelihood estimate for a given cell.
    Initialize $m_{i,j,k} = 1 \forall i, j, k$
    for $c \in \{0, MAX\_ITERATIONS\}$ do
        for $i, j, k \forall$ <Event-Type, Population, Country> do
            $\hat{m}^{c+1} = \hat{m}^{c}_{ijk} * \frac{x_{ij+}}{\hat{m}^{c}_{ij+}}$
            $\hat{m}^{c+2} = \hat{m}^{c+1}_{ijk} * \frac{x_{i+k}}{\hat{m}^{c}_{i+k}}$
            $\hat{m}^{c+3} = \hat{m}^{c+2}_{ijk} * \frac{x_{+jk}}{\hat{m}^{c}_{+jk}}$
        if $\hat{m}^{c} - \hat{m}^{c-1} < \tau$ then
      return $\hat{m}^{c}$
      return $\hat{m}^{c}$
```

Fig. 19 summarizes these uncertainties.

Among all the incidents of civil unrest that we encounter, the largest and the most significant ones are planned events. These events are usually organized by political parties, labor and student unions. Since it takes a huge effort to organize protest demonstrations that attract thousands, the organizers must disseminate information regarding the venue and the date and time. These announcements are posted on the organizers' websites and are widely shared on social media. By scouring our sources, it has been possible for EMBERS to accurately forecast the occurrence of these types of protests.

The recurring events take place on a regular basis. For instance, in Chile and Argentina the 'mothers of the disappeared' protest the disappearance of their children by the military dictatorships of the 1970's and 19780's on a certain particular day of the week and in the same plaza. In some countries with large Muslim populations, fighting and protests break out regularly after Friday evening prayer as people stream out of the mosques after listening to fiery sermons. These are typically small events but if they are reported as part of our GSR, EMBERS models will be able to forecast them.

The protests for which the causes are known but not the timing are staged spontaneously. These events are the outcomes of longstanding frustration and anger which fuel widespread protests in response to trigger events. Thus, the viral videos of police brutality or a sudden change in government policy can start a prairie fire of protests. The Brazilian Spring with origins in bus fares and which channeled public anger against corruption and government mismanagement is a classical example. The challenge here is not just to be aware of the underlying tensions that might erupt when an event occurs, but to also distinguish between events that do and do not perform as triggers. Algorithms to better model precursors is an area of further research that will aid further forecasting this class of events.

Finally, *black swan* events [18] are rare and truly unforeseen and can happen as a result of natural disasters, the sudden death of a leader, or even the sudden rise of a small group that can truly destabilize a nation. For instance the rise of the Islamic State in Iraq and Syria (ISIS) has truly confounded policymakers all over the world. While there were other Sunni groups, from al-Qaeda to al-Nusra, that contributed to instability, the rapid ascendance of ISIS, which did not depend on an isolated terrorist attack and burst out with a clear holding of territories as a full-scale insurgency, surprised most observers. It might not be feasible to forecast the beginnings of such events; however, once such movements have been initiated, models should be able to detect and forecast their momentum.

8 Ethical Issues

EMBERS, as an anticipatory intelligence system, has many powerful legitimate uses but is also susceptible to abuse.

First, it is important to have a discussion of civil unrest and its role in society. In the proper circumstances, civil unrest enhances the ability of citizens to communicate not only their views but also their priorities to those who govern them. Governments constantly need to make choices and find it difficult to know, on specific issues at particular times, how their constituencies value the available options. Elections are retrospective indicators and rarely issue-specific; polling taps into sentiment, but is not a good indicator of priorities or strength of feeling because of the low cost associated with responding. Events, on the other hand, indicate a willingness to bear some costs (organization, mobilization, identification) in support of an issue and thus reveal not only preferences but provide some indication of priorities.

An open sources indicators approach, as used here, is a potentially powerful tool for understanding the social construction of meaning and its translation into behavior. EMBERS can contribute to making the transmission of citizen preferences to government less costly to the economy and society as well. There are economic costs to even peaceful disruptions embodied in civil unrest due to lost work hours and the deployment of police to manage traffic and the interactions between protestors and bystanders. Given the vulnerability of large gatherings to provocation by handfuls of violence-oriented protestors (e.g., Black Box anarchists in Brazil) the economic, social and political costs of large-scale public demonstrations are also potentially significant to marchers, bystanders, property owners and the government – democratically elected or not. The right to demonstrate can still be respected but if the government responds to grievances in time, the protestors may cancel the event or fewer people might participate in the event. In today's interconnected society, protests also cause disruptions to supply chain logistics, travel, and other sectors, and anticipating disruptions is key to ensuring safety as well as reliability.

The potential power of civil unrest forecasting systems, like those of most scientific advances, is susceptible to abuse by both democratic and non-democratic governments.

The appropriate safeguards require developing transparent and accountable democratic systems, not outlawing science. Non-democratic governments may clearly abuse such forecasting systems. But even here the value of forecasting civil unrest is not simply negative. Many non-democratic regimes transition to democratic ones, often in a violent process but not always (in Latin America, authoritarian regimes negotiated transitions to democracy without a civil war in Mexico, Honduras, Peru, Bolivia, Brazil, Uruguay, Argentina and Chile). The rational choice models of authoritarian decision-making in such crises always explain a dictatorships' collapse rather than accommodation to a transition by pointing to the lack of credible information in a dictatorship regarding citizens' true feelings. EMBERS-like models may thus provide the information that facilitates and encourages some transitions.

9 Conclusions

We have presented the successes as well as the lessons learned from the EMBERS architecture as a result of four years of 24x7 operation. EMBERS has shown itself to be a reliable predictor of civil unrest events in 10 different countries, and three different major languages. Yet despite these successes EMBERS has also been a learning experience and even in its misses much has been learned.

Future work falls primarily in two directions.. First, we will continue to build a strong capability to forecasting societal events, leveraging multiple, overlapping, models of selective superiority. Being able to better understand how multiple alerts from different models should be fused would be an invaluable tool to analysts. Secondly, we are investigating techniques to remove or reduce the human element required in generating the GSR. Currently the most human intensive part of the EMBERS project is generating a GSR for training and validation of the models.

References

[1] Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice.* Springer Science & Business Media, 2007.

[2] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan, and A. Vullikanti. Forecasting Social Unrest Using Activity Cascades. *PLoS One*, 10(6):e0128879, 2015.

[3] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, et al. Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions. In *Proc. SIAM Int. Conf. Data Min. (SDM 2014)*, Philadelphia, PA, 2014.

[4] R. Dingledine, N. Mathewson, and P. Syverson. Tor: the second-generation onion router. page 21, aug 2004.

[5] A. Doyle, G. Katz, K. Summers, C. Ackermann, et al. Forecasting Significant Societal Events Using The Embers Streaming Predictive Analytics System. *Big Data*, 2(4):185–195, dec 2014.

[6] J. A. Goldstone, R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward. A Global Model for Forecasting Political Instability. *Am. J. Pol. Sci.*, 54(1):190–208, jan 2010.

[7] A. Hoegh, S. Leman, P. Saraf, and N. Ramakrishnan. Bayesian Model Fusion for Forecasting Civil Unrest. *Technometrics*, 57(3):332–340, feb 2015.

[8] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan. Detecting and forecasting domestic political crises. In *Proc. 2014 ACM Conf. Web Sci. - WebSci '14*, pages 192–196, New York, New York, USA, jun 2014. ACM Press.

[9] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan. Combining Heterogeneous Data Sources for Civil Unrest Forecasting. In *Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM '15*, pages 258–265, New York, New York, USA, aug 2015. ACM Press.

[10] K. Leetaru and P. Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. *ISA Annual Convention*, pages 1979–2012, 2013.

[11] H. Llorens, L. Derczynski, R. J. Gaizauskas, and E. Saquete. Timen: An open temporal expression normalisation resource. In *Proceedings of Language Resources and Evaluation*, LREC, pages 3044–3051, 2012.

[12] A. Mahendiran, W. Wang, S. Lira, J. Arredondo, B. Huang, L. Getoor, D. Mares, and N. Ramakrishnan. Discovering evolving political vocabulary in social media. In *Behavior, Economic and Social Computing (BESC), 2014 International Conference on*, pages 1–7. IEEE, 2014.

[13] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan. Planned protest modeling in news and social media. 2015.

[14] S. P. O'Brien. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *Int. Stud. Rev.*, 12(1):87–104, Mar. 2010.

[15] N. Ramakrishnan, P. Butler, S. Muthiah, et al. 'beating the news' with embers: Forecasting civil unrest using open source indicators. KDD '14, pages 1799–1808, New York, NY, USA, 2014. ACM.

[16] T. Rekatsinas, S. Ghosh, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, and N. Ramakrishnan. Sourceseer: Forecasting rare disease outbreaks using multiple data sources. *Timeline*, 7(8), 2015.

[17] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[18] N. N. Taleb. *The Black Swan. The Impact of the Highly Improbable.* Random House Inc., 2008.

[19] M. D. Ward, N. W. Metternich, C. Carrington, C. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle. Geographical models of crises: Evidence from icews. *Advances in Design for Cross-Cultural Activities*, page 429, 2012.

[20] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan. Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling. *PLoS ONE*, 9(10):e110206, 2014.