Capturing Planned Protests from Open Source Indicators

Sathappan Muthiah, Bert Huang, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz, Naren Ramakrishnan

■ *Civil unrest events (protests, strikes, and "occupy" events) are common occurrences in both democracies and authoritarian regimes. The study of civil unrest is a key topic for political scientists as it helps capture an important mechanism by which citizens express themselves. In countries where civil unrest is lawful, qualitative analysis has revealed that more than 75 percent of the protests are planned, organized, or announced in advance; therefore detecting references to future planned events in relevant news and social media is a direct way to develop a protest forecasting system. We report on a system for doing that in this article. It uses a combination of key-phrase learning to identify what to look for, probabilistic soft logic to reason about location occurrences in extracted results, and time normalization to resolve future time mentions. We illustrate the application of our system to 10 countries in Latin America: Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. Results demonstrate our successes in capturing significant societal unrest in these countries with an average lead time of 4.08 days. We also study the selective superiorities of news media versus social media (Twitter, Facebook) to identify relevant trade-offs.*

Civil unrest events (protests, strikes, and "occupy" events) are common occurrences in both democracies and in authoritarian regimes. Although we typically associate civil unrest with disruptions and instability, social scientists believe that civil unrest reflects the democratic process by which citizens communicate their views and preferences to those in authority. The advent of social media has afforded the citizenry new mechanisms for organization and mobilization, and online news sources and social networking sites like Facebook and Twitter can provide a window into civil unrest happenings in a particular country (see figure 1).

 ISSN 0738-4602

20 de enero 2016

TalCual
claro y raspao

OPINIÓN NACIÓN ECONOMÍA EDITORIAL MUNDO OTRAS HISTORIAS

TEMAS DEL DIA ASAMBLEA NACIONAL | TSJ | DIVINA PASTORA

INICIO NACIÓN

Que la calle no calle

28-02-2014

ANA MARÍA LÓPEZ

A pesar de que el Gobierno insiste en promulgar la paz la concentración de ayer terminó con gases lacrimógenos. La GN volvió a salirse con las suyas y haciendo usos de las ballenas reprimieron otra manifestación pacífica, sin embargo, los estudiantes no se dan por vencidos y anunciaron que marcharán el domingo

La concentración convocada por el movimiento estudiantil en Caracas no culminó pacíficamente. Aunque desde las 11 de la mañana hasta las 2 de la tarde todo transcurrió con normalidad, a eso de las 2:30 pm, cuando la mayoría de los que se encontraban en la avenida Venezuela de El Rosal se disponían a irse, otros decidieron trasladarse hasta la autopista Francisco Fajardo para trancarla.

Fue en ese momento cuando efectivos de la Guardia Nacional accionaron sus bombas lacrimógenas contra los manifestantes para impedir que realizaran la toma.

Después la arremetida, a través de su cuenta twitter Juan Requesens, presidente de la Federación de Centros de Estudiantes de la Universidad Central de Venezuela (FCU-UCV), criticó que se hable de paz y luego se utilicen acciones violentas por parte de las fuerzas de seguridad: "Hablan de paz y después que los estudiantes nos concentramos pacíficamente gritando Ni un muerto más, nos lanzan bombas lacrimógenas".

El alcalde de Baruta, Gerardo Blyde, consideŕo que fue "excesiva" la represión de la GN hacia los manifestantes en Las Mercedes. Pasadas las 4 de la tarde la arremetida contra los jóvenes continuó, esta vez desde la Plaza Altamira en Chacao.

El próximo domingo los universitarios esperan mantener la actividad de calle. Es por ello que convocaron a una marcha en la capital, donde esperan congregar a ciudadanos de todos los sectores que saldrán desde distintos puntos a la Plaza Brión, en Chacaíto.

En las próximas horas deben confirmar ruta. "No nos arrodillamos seguiremos exigiendo justicia, igualdad y paz. Luchamos con el pueblo por sus derechos. Escribió Requesens.

Figure 1. An Example Article Describing Plans for a Future Protest (Venezuela, June 11, 2014)

Why Study and Forecast Protests?

Our primary region of interest is Latin America (secondary regions of interest include the Middle East and North Africa), and protest is an important topic of study in this region, as many countries here are democracies struggling to consolidate themselves. The combination of weak channels of communication between citizen and government, and a citizenry that still has not grasped the desirability of elections as the means to affect politics means that public protest will be an especially attractive option. To illustrate the power of protest in Latin America we need only recall that between 1985 and 2011, 17 presidents resigned or were impeached under pressure from demonstrations, usually violent, in the streets. Protests have also resulted in the rollback of price increases for public services, such as during the Brazilian Spring of June 2013.

Forecasting protests is an important capability in many domains. For the tourism industry, forecasting protests can support the issuance of travel warnings. For law enforcement, forecasting protests can aid in preparedness. For social scientists, forecasting protests will provide insight into how citizens express themselves. For governments, a protest forecasting system can help prioritize citizen grievances. Finally, protests can have a debilitating effect on multiple industries (especially those that rely on worldwide supply chain management) and thus a protest forecasting system can aid in planning and design of alternative travel and shipping routes.

Planned Protests

Our basic hypothesis is that protests that are larger will be more disruptive and communicate support for their cause better than smaller protests. Mobilizing large numbers of people is more likely to occur if a protest is organized and the time and place announced in advance. Because protests are costly and more likely to succeed if they are large, we should expect planned, rather than spontaneous, protests to be the norm. Indeed, in a sample of 288 events from our study selected for qualitative review of their antecedents, for 225 we located communications regarding the upcoming occurrence of the event in media, and only 49 could be classified as spontaneous (we could not determine whether communications had or had not occurred in the remaining 14 cases).

EMBERS

We are an industry-university partnership charged with developing an automatic protest forecasting system for 10 countries in Latin America (LA) and 7 countries in the Middle East and North Africa region (MENA). Our system, called EMBERS (for early model-based event recognition using surrogates), has been deployed since November 2012 and has been generating forecasts (called warnings or alerts) automatically, without a human in the loop, since then (the MENA region was added in July 2014). These forecasts are emailed to a third party (the MITRE Corporation) for evaluation. Analysts at Mitre organize a reference database of protests (called the gold standard report, or GSR) by surveying newspapers for reports of protests, and compare our warnings against the GSR to generate a scoring report, using evaluation criteria described later.

The full EMBERS system has been described elsewhere (Ramakrishnan et al. 2014; Doyle et al. 2014), including the overall system architecture, data sources used for analysis, and the various forecasting models that make up the system. EMBERS adopts a multimodel approach, wherein different models are leveraged for their selective superiorities to generate a fused set of alerts. One of the best performing models in EMBERS is the planned protest model that detects ongoing organizational activity and generates warnings accordingly. This article is the first to present this model in detail, including the research issues involved, and how we addressed them in EMBERS.

Capturing mentions of protest planning and organization is not as easy as it might appear. First, articles of interest are written in different languages (Spanish, Portuguese, French, Dutch, and English). Second, multiple locations are often mentioned in a given article, leading to (natural language) ambiguity about the intended location of the event. Significant reasoning is required to discern the correct protest location. Finally, identifying the date of a planned event in Latin America involves significant multilingual temporal-semantic natural language processing as dates are often described with vague, relative, or otherwise context-dependent expressions (for example, Sunday or in two days).

Our detection approach combines shallow linguistic analysis to identify relevant documents (articles, tweets) with targeted deep semantic analysis of the selected documents. Despite its simplicity, this approach is capable of detecting indicators of event planning with surprisingly high accuracy.

Our contributions follow:

First, we present a protest forecasting system that couples three key technical ideas: key-phrase learning to identify what to look for, probabilistic soft logic to reason about location occurrences in extracted results, and date normalization to resolve future tense mentions. We demonstrate how the integration of these ideas achieves objectives in precision, recall, and quality (accuracy).

Second, we illustrate the application of our system to 10 countries in Latin America: Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. Our system predicts the when of the protest as well as the where of the protest (down to a city-level granularity). We conduct ablation studies to identify the relative contributions of news media (news + blogs) versus social media (Twitter, Facebook) to identify future happenings of civil unrest. Through these studies we illustrate the selective superiorities of different sources for specific countries.

Third, unlike many studies of retrospective forecasting of protests, our system has been deployed and in operation for nearly three years. The end consumers of our alerts are analysts studying Latin America. Our results demonstrate that we are able to capture significant societal unrest in the above countries with an average lead time of 4.08 days.

Related Work

Five categories of related work are briefly discussed here (see also table 1): (1) event detection through text extractions; (2) temporal information extraction; (3) event forecasting; (4) future retrieval; and (5) planned protest detection. These categories are discussed briefly in the following paragraphs.

Event detection through text extractions is an extensively studied topic in the literature. Document-clustering techniques are used in, for example, Gabrilovich, Dumais, and Horvitz (2004) to identify events retrospectively or as the stories arrive. Work such as that of Banko et al. (2007), Chambers and Jurafsky (2011), and Riloff and Wiebe (2003) focuses on extraction patterns (templates) to extract information from text.

Temporal information extraction is another well-studied topic. The TempEval challenge (Verhagen et al. 2009) led to a significant amount of algorithmic development for temporal natural language processing (NLP). For instance a specification language for temporal and event expressions in natural language text is described in Pustejovsky et al. (2003). References such as Llorens et al. (2012) and Mani and Wilson (2000) provide methods to resolve temporal expressions in text (our own work here uses the TIMEN package [Llorens et al. 2012] and more recently the HeidelTime package [Strötgen et al. 2014]).

Under the event forecasting category, Radinsky and Horvitz (2013) find event sequences from a corpora and then use these sequences to determine whether an event of interest (for example, a disease outbreak, or a riot) will occur sometime in the future. This work predicts only whether a potential event will happen given a historical event sequence but does not geolocate the event to a city-level resolution, as we do here. Kallus (2014) makes use of event data from RecordedFuture (Truvé 2011) to determine whether a significant protest event will occur in the subsequent three days and casts this as a classification problem. This work only focuses on prediction of significant events (suitably defined) and the forecast is limited to a fixed time window of the next three days.

Baeza-Yates (2005) provides one of the earliest discussions of the future retrieval topic; here future temporal information in text is found and used to retrieve content from search queries that combine both text and time with a simple ranking scheme. Kawai et al. (2010) present a search engine (ChronoSeeker) for searching future and past events. RecordedFuture (Truvé 2011), introduced earlier, conducts real-time analysis of news and tweets to identify mentions of events along with associated times. Anecdotally it is estimated that approximately (only)

	Relative Date Resolution	Ingest Multiple Sources?	Reasoning About Location	Learning Word/Phrase Filters
Future Search Engines (Kawai et al. 2010; Jatowt and Au Yeung 2011; Baeza-Yates 2005)	✓			
Time-to-Event Recognition (Tops, van den Bosch, and Kunneman 2013; Hurriyetoglu, Kunneman, and van den Bosch 2013)	✓			
Planned Protest Detection (Xu et al. 2014; Compton et al. 2013)		✓		
This article	✓	✓	✓	✓

Table 1. Comparison of Our Approach Against Other Techniques.

ENTITY(*L*, *location*) ∧ REFERSTO(*L*, *locID*) → PSLLOCATION(*Article*, *locID*)

Rule 1.

5–7 percent of events extracted by RecordedFuture are about the future.

In the planned protest detection category, two publications — Compton et al. (2013) and Xu et al. (2014) — align very closely to our own work as their emphasis is on protest forecasting. Both works are aimed at forecasting protests but emphasize different data sources and different methodologies. For instance, the work in Compton et al. (2013) filters the Twitter stream for key words of interest and searches for future date mentions in only absolute terms, that is, explicit mentions of a month name and a number (date) less than 31. Such an approach will not be capable of extracting the more common way in which future dates are referenced, for example, phrases like "tomorrow," "next Tuesday." The work by Xu et al. (2014) by the same group of authors uses the Tumblr feed with a smaller set of key words but again is restricted to the use of absolute time identifiers.

Probabilistic Soft Logic

In this section, we briefly describe probabilistic soft logic (PSL) (Kimmig et al. 2012), a key component of our geocoding strategy. PSL is a framework for collective probabilistic reasoning on relational domains. PSL represents the domain of interest as logical atoms. It uses first-order logic rules to capture the dependency structure of the domain, based on which it builds a joint probabilistic model over all atoms. Instead of hard truth values of 0 (false) and 1 (true), PSL uses soft truth values relaxing the truth values to the interval [0,1]. The logical connectives are adapted accordingly.

User-defined predicates are used to encode the relationships and attributes and rules capture the dependencies and constraints. The rules can also be labeled with nonnegative weights, which are used during the inference process. The set of predicates and weighted rules thus make up a PSL program where known truth values of ground atoms are derived from observed data and unknown truth values for the remaining atoms are learned using the PSL inference.

Example One

We will follow a running example throughout this section. In our geocoding subtask, we create a PSL program that reasons about the predicate REFERSTO, which maps a text string to real-world locations. For example, REFERSTO("DC," washingtonDC) evaluates to true if the knowledge base believes that the article refers to Washington D.C. as "DC." This predicate gets used in rules that define dependencies between predicates, such as that shown in rule 1.

Rule 1 states that an entity extracted from an article text that matches a known REFERSTO mapping implies that the PSL program's predicted location will follow that mapping. Some of these logical atoms will be known as parts of a knowledge base, while others will be unknown and will be inferred by PSL.

Given a set of atoms $\ell = \{\ell_1, \ldots, \ell_n\}$, an interpretation defined as $I : \ell \rightarrow [0, 1]^n$ is a mapping from atoms to soft truth values. PSL defines a probability distribution over all such interpretations such that those that satisfy more ground rules are more probable. Lukasiewicz *t*-norm and its corresponding co-norm are used for defining relaxations of the logical AND and OR, respectively, to determine the degree to which a ground rule is satisfied. Given an interpretation *I,* PSL defines the formulas for the relaxation of the logical conjunction (∧), disjunction (∨), and negation (¬) as follows:

$$\ell_1 \wedge \ell_2 = \max\{0, I(\ell_1) + I(\ell_2) - 1\},$$
$$\ell_1 \vee \ell_2 = \min\{I(\ell_1) + I(\ell_2), 1\},$$
$$\neg l_1 = 1 - I(\ell_1),$$

The interpretation *I* determines whether the rules are

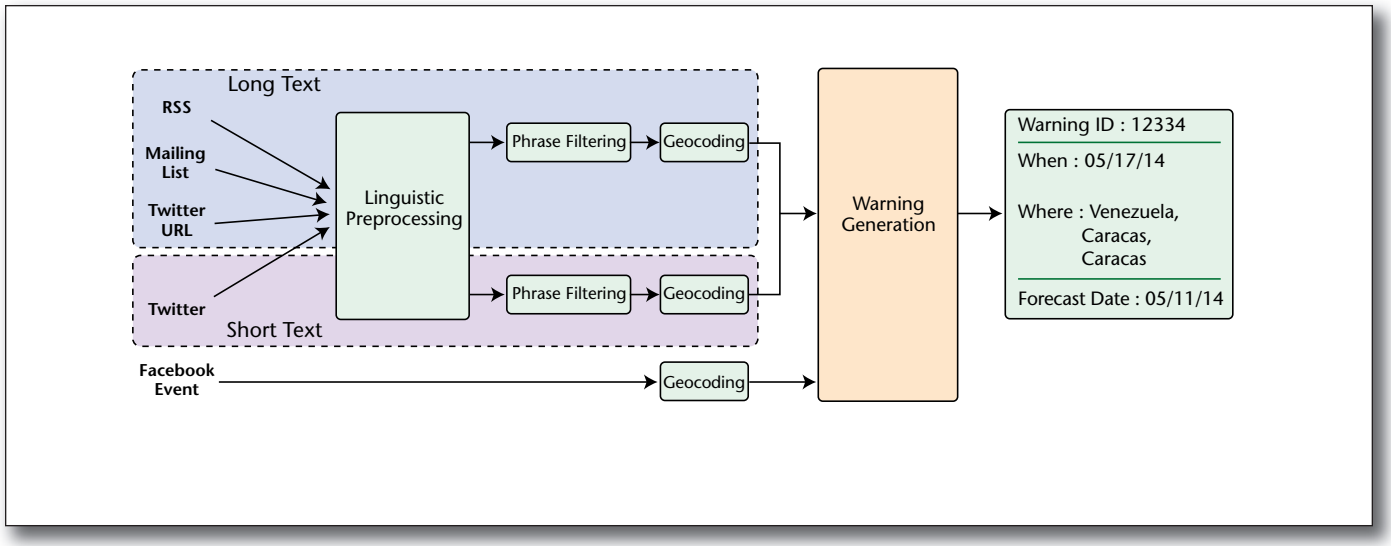

Figure 2. Schematic of the Planned Protest Detector That Ingests Five Different Types of Data Sources.

satisfied. A rule $r \equiv r_{body} \rightarrow r_{head}$ is satisfied if and only if the truth value of the head is at least that of the body. Otherwise, PSL uses a distance to satisfaction, which measures the degree to which this condition is violated

$$d_r(I) = \max\{0, I(r_{body}) - I(r_{head})\}.$$

Example Two

Continuing the previous example, if we have the rule shown in rule 2 and the truth values of the known atoms are whoen in value 1, then following the relaxation of logical AND, the truth value of the antecedent is max{0, 1.0 + 0.6 – 1} = 0.6. Moreover, on one hand, if the truth value of the atom to be inferred, PSLLocation(*Article, washingtonDC*), is 0.1, then the distance to satisfaction of this rule is max{0, 0.6 – 0.1} = 0.5. On the other hand, if the truth value of the head is 0.7, then the distance to satisfaction is max{0, 0.6 – 0.7} = 0, meaning the rule is satisfied.

PSL then induces a probability distribution over possible interpretations I over the given set of ground atoms ℓ in the domain. If R is the set of all ground rules that are instances of a rule from the system and uses only the atoms in I then, the probability density function f over I is defined as

$$f(I) = \frac{1}{Z}\exp\left(-\sum_{r\in R}\lambda_r(d_r(I))^p\right) \quad (1)$$

$$Z = \int_I \exp\left(-\sum_{r\in R}\lambda_r(d_r(I))^p\right) \quad (2)$$

where λ_r is the weight of the rule r, Z is a normalization constant, and $p \in 1, 2$ provides a choice between linear or quadratic loss functions, which produce different modeling behavior. PSL further allows inclusion of linear equality and inequality constraints, which enable modeling of functional constraints on the domains and ranges of predicates.

Entity(*"Washington," location*) ∧ RefersTo(*"Washington," washingtonDC*) → PSLLocation(*Article, washingtonDC*)

Rule 2.

Entity(*"Washington," location*) : 1.0
RefersTo(*"Washington,"washingtonDC*) : 0.6

Value 1.

The probability distribution in equation 1 is an example of a hinge-loss Markov random field (Bach et al. 2013), which has a form of energy function that makes inference of the most probable explanation an efficient convex optimization. The expressiveness of PSL and the efficiency of inference in its models allows us to encode dependencies between various aspects of geolocation that are jointly inferred.

Example Three

The joint probability distribution enables PSL to reason about conflicting evidence, for example if we additionally have the atom RefersTo(*"Washington," Washington State*) in our knowledge base with truth value 0.2, we would have two conflicting PSL Location implications. The probabilistic, weighted rules, as well as the soft truth values of known atoms control how much PSL considers each piece of evidence, and additional corroborating or conflicting evidence would also be incorporated into the final joint inference.

Approach

The general approach we adopt is to identify open source documents that appear to indicate civil unrest event planning, extract relevant information from identified documents, and use that as the basis for a structured warning about the planned event. Each of these processing steps (see figure 2) is outlined next.

Linguistic Preprocessing

Since our region of interest is Latin America as well as the Middle East and North Africa, the collection of text harvested is inherently multilingual, with Spanish, Portuguese, Arabic, and English as the dominating languages. Ingested documents are subjected to shallow linguistic processing prior to analysis. This initial processing involves identifying the language of the document, distinguishing the words (tokenization), normalizing words for inflection (lemmatization), and identifying expressions referring to people, places, dates, and other entities. We use Basis Technology's Rosette Linguistics Platform (RLP) suite of multilingual commercial tools[1] for this processing. The output of this linguistic preprocessing serves as input to subsequent deeper analysis in which date expressions are normalized, sentences that appear to be describing protest planning are identified, and the geographic focus of the text computed.

Date Normalization

Date processing is particularly crucial to the identification and interpretation of statements about future events. We used the TIMEN (Llorens et al. 2012) date normalization package to normalize date expressions in English, Spanish, and Portuguese (we extended the system to cover Portuguese) and HeidelTime (Strötgen et al. 2014) to do the same for Arabic (extending the set of rules to handle Hijri dates). Both systems' rules were extended to improve coverage and accuracy on our document collection.

The systems make use of metadata such as the day of publication and other information about the linguistic context to determine for each date expression, what day (or week, month, or year) it refers to. For example in a tweet produced on June 10, 2014, the occurrence of the term *Friday* used in a future-tense sentence, "We'll get together on Friday," will be interpreted as June 13, 2014. Each expression identified as a date by the RLP preprocessor is normalized in this way, with accuracy of just over 80 percent on our data set.

Phrase Filtering

In order to identify relevant documents, input documents are filtered on a set of key phrases, that is, the text of the document is searched for the presence of one or more key phrases in a list of phrases that are indicative of an article's focus being a planned civil unrest event. The list of key phrases indicating civil unrest planning was obtained in a semiautomatic manner, as detailed next. Articles that do match are processed further, those that do not are ignored.

Phrase Matching

Our key-phrase matching is highly general and linguistically sophisticated. The phrases in our list are general rules for matching, rather than literal string sequences. Typically a phrase specification comprises: two or more word lemmas, a language specification, and a separation threshold. This indicates that words — potentially inflected forms — in a given sequence potentially separated by one or more other words, should be taken to be a match. We determined that this kind of multiword key phrases was more accurate than simple key words for extracting events of interest from the data stream.

The presence of a key phrase is checked by searching for the presence of individual lemmas of the key phrase within the same sentence separated by at most a number of words that is fewer than the separation threshold. This method allows for linguistically sophisticated and flexible matching, so, for example, the key phrase [*plan protest,* 4, English] would match the sentence The students are planning a couple big protests tomorrow in an input document.

Phrase List Development

The set of key phrases was tailored (slightly) to the genre of the input. In particular different phrases were used to identify relevant news articles and blogs from those used to filter tweets. The lists themselves were generated semiautomatically.

Initially, a few seed phrases were obtained manually with the help of subject matter experts. An analysis of news reports for planned protests in the print media helped create a minimum set of words to use in the query. We choose four nouns from the basic query that is used predominantly to indicate civil unrest in the print media — demonstration, march, protest, and strike. We translated them into Spanish and Portuguese, including synonyms. We then combined these with future-oriented verbs, for example, to organize, to prepare, to plan, and to announce. For Twitter, shorter phrases were identified, and these had a more direct call for action, for example, marchar, manhã de mobilização, vamos protestar, huelga.

To generalize this set of phrases, the phrases were then parsed using a dependency parser (Padró and Stanilovsky 2012) and the grammatical relationship between the core nominal focus word (for example, protest, manifestación, huelga) and any accompanying word (for example, plan, call, anunciar) was extracted. These grammatical relations were used as extraction patterns as in the paper by Riloff and Wiebe (2003) to learn more phrases from a corpora of sentences extracted from the data stream of interest (either news/blogs or tweets). This corpus consists of sentences that contained any one of the nominal focus words and also had mentions of a future date. The separation threshold for a phrase was also learned, being set to the average number of

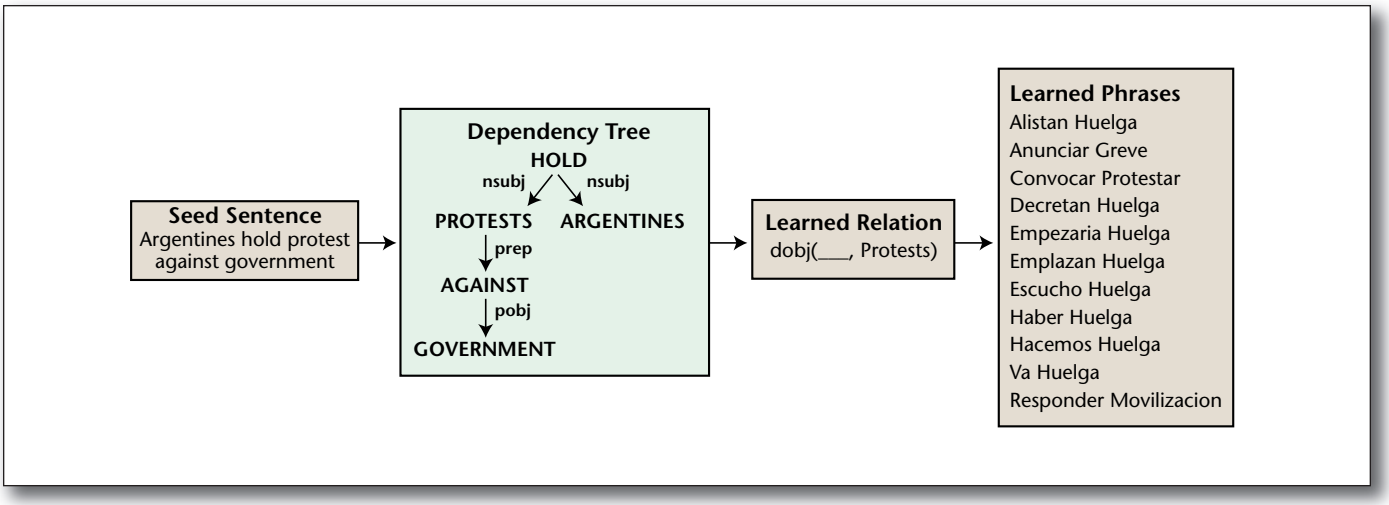

Figure 3. An Example of Phrase Learning for Detecting Planned Protests.

ENTITY(*L*, location) ∧ REFERSTO(*L*, *locID*) → PSLLOCATION(Article, locID)

ENTITY(*C*, location) ∧ ISCOUNTRY (C) → ARTICLECOUNTRY(Article, *C*)

ENTITY(*S*, location) ∧ ISSTATE(*S*) → ARTICLESTATE(Article, *S*)

Rule 3.

words separating the nominal focus and the accompanying word.

The set of learned phrases is then reviewed by a subject matter expert for quality control. Using this approach, we learned 112 phrases for news articles and blogs and 156 for tweets. This phrase-learning process is illustrated in figure 3.

Geocoding

After linguistic preprocessing and suitable phrase filtering, messages are geocoded with a specification of the geographical focus of the text — specified as a city, state, country triple — that indicates the locality that the text is about. We make use of different geocoding methodologies for Twitter messages, for Facebook Events pages, and for news articles and blogs. These are described below.

Twitter and Facebook

For tweets, the geographic focus of the message is generated by a fairly simple set of heuristics involving (1) the most reliable but least available source, that is, the geotag (latitude, longitude) of the tweet itself, (2) Twitter places metadata, and (3) if the aforementioned are not available, the text fields contained in the user profile (location, description) as well as the tweet text itself to find mentions of relevant locations. Additional toponym disambiguation heuristics are used to identify the actual referent of the mention. Similar methods are used to geocode event data extracted from Facebook event pages.

News and Blogs

For longer articles such as news articles, the geographic focus of the message is identified using much more complex methods to extract the protest location from news articles. We use PSL to build a probabilistic model that infers the intended location of a protest by weighing evidence coming from entities extracted by the RLP preprocessor and information in the *World Gazetteer.*

The primary rules in the model encode the effect that RLP-extracted location strings that match to gazetteer aliases are indicators of the article's location, whether they be country, state, or city aliases. Each of these implications is conjuncted with a prior probability for ambiguous, overloaded aliases that are proportional to the population of the gazetteer location. For example, if the string "Los Angeles" appears in the article, it could refer to either Los Angeles, California, or Los Ángeles in Argentina or Chile. Given no other information, our model would infer a higher truth value for the article referring to Los Angeles, California, because it has a much higher population than the other options (rule 3).

Note that these are not deterministic rules; for example, they do not use the logical conjunction but rather the Lukasiewicz *t*-norm based relaxation. Further, these rules do not fire deterministically but are instead simultaneously solved for satisfying assignments as described in the section on probabilistic soft logic.

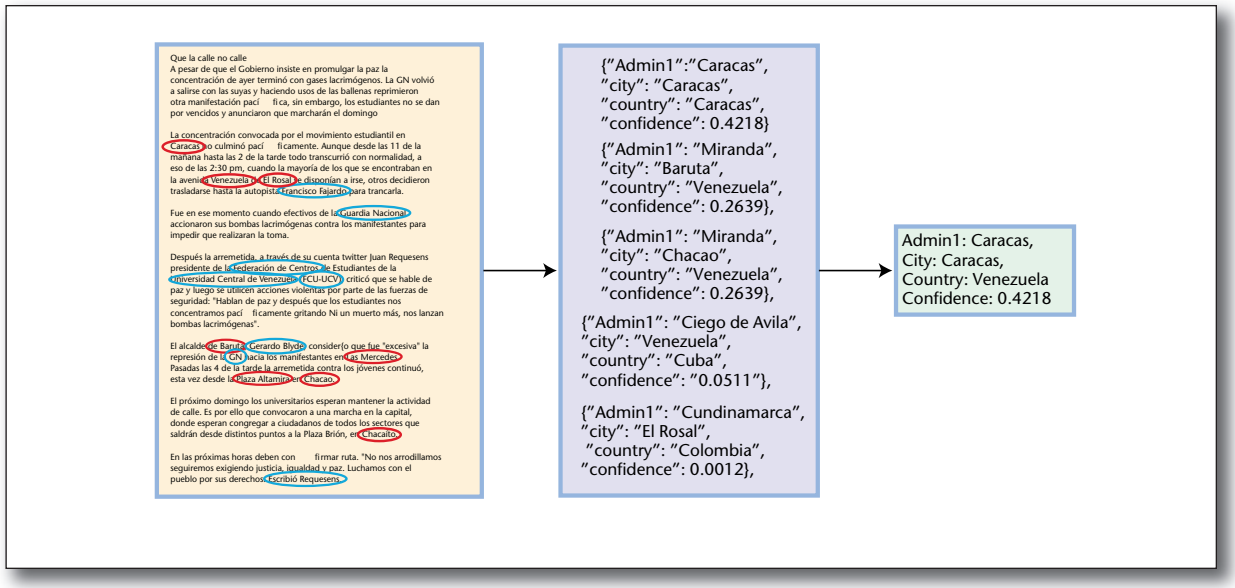

Figure 4. An Example of Location Inference Using PSL.

Red (dark gray) circles denote named entities identified as locations and blue (light gray) denotes other types of entities. The article describes students planning a march on Sunday. It identifies multiple locations, for example, Chacao, El Roso, and the Francisco Fajardo highway where protests have been happening. There is also a reference to a quote by the mayor of Baruto. Mentions of such multiple locations are resolved using our PSL program to the intended location, here Caracas.

ENTITY(*O*, organization) ∧ REFERSTO(*O*, locID) → PSLLOCATION(Article, locID)

ENTITY(*O*, organization) ∧ ISCOUNTRY(*O*) → ARTICLECOUNTRY(Article, *O*)

ENTITY(*O*, organization) ∧ ISSTATE(*O*) → ARTICLESTATE(Article, *O*)

Rule 4.

PSLLOCATION(Article, locID) ∧ COUNTRY(locID, *C*) → ARTICLECOUNTRY(Article, *C*)

PSLLOCATION(Article, locID) ∧ ADMIN1(locID, *S*) → ARTICLESTATE(Article, *S*)

Rule 5.

The secondary rules, which are given half the weight of the primary rules, perform the same mapping of extracted strings to gazetteer aliases, but for extracted persons and organizations. Strings describing persons and organizations often include location clues (for example, "mayor of Buenos Aires"), but intuition suggests the correlation between the article's location and these clues may be lower than with location strings (rule 4).

Finally, the model includes rules and constraints to require consistency between the different levels of geolocation, making the model place higher probability on states with its city contained in its state, which is contained in its country. As a postprocessing step, we enforce this consistency explicitly by using the inferred city and its enclosing state and country, but adding these rules into the model make the probabilistic inference prefer consistent predictions, enabling it to combine evidence at all levels (rule 5).

As an example of how PSL aids in location identification, the protest from figure 1 is revisited in figure 4, which illustrates the evidence that the PSL model gathers from the news article and the inferred locations.

We evaluated our planned protest detection system for Latin America using metrics similar to those described in Ramakrishnan et al. (2014). Given a set of alerts issued by the system and the GSR comprising actual protest incidents, we aim to identify a correspondence between the two sets through a bipartite matching. An alert can be matched to a GSR event only if (1) they are both issued for the same country, (2) the alert's predicted location and the

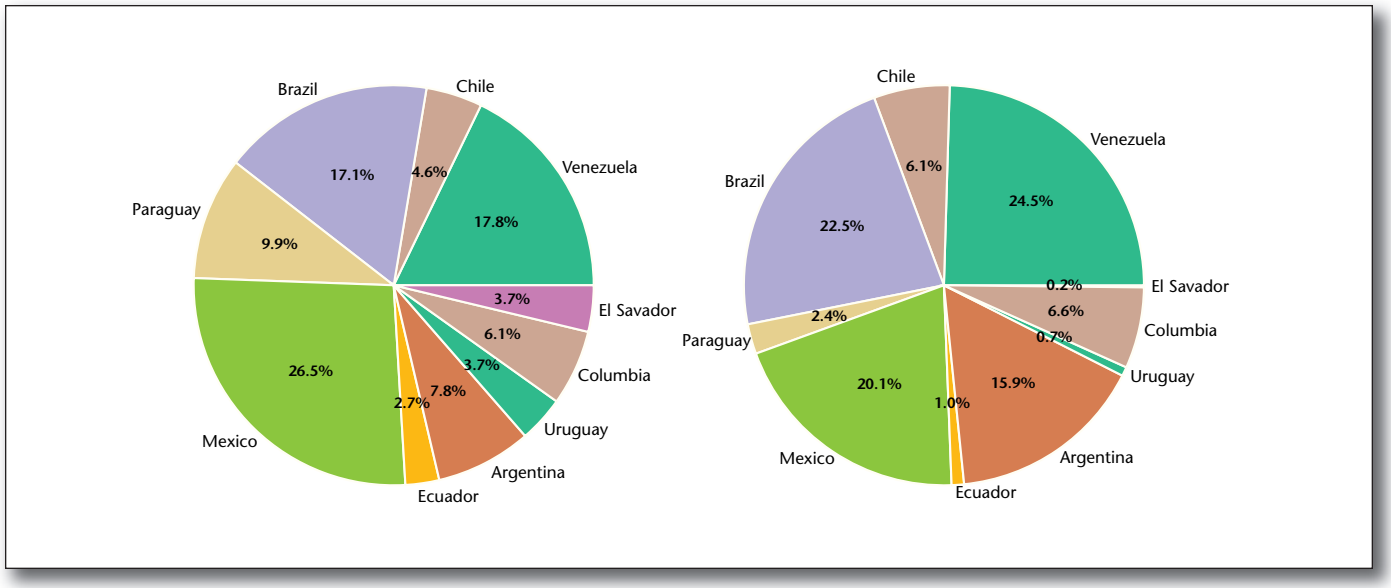

Figure 5. Distribution of Alerts and GSR Events Across the Latin American Countries Studied in This Article.

	News / Blogs				Twitter				Facebook				Combined			
	QS	Pr.	Rec.	LT	QS	Pr.	Rec.	LT	QS	Pr.	Rec.	LT	QS	Pr.	Rec.	LT
AR	3.14	0.32	0.69	3.94	3.52	0.78	0.14	3.14	3.70	0.50	0.04	3.00	3.02	0.36	0.80	4.50
BR	3.14	0.48	0.54	5.85	–	–	–	–	3.62	0.76	0.18	2.46	3.28	0.49	0.65	5.15
CL	3.06	0.91	0.67	5.40	3.52	1.00	0.23	4.29	–	–	–	–	3.16	0.83	0.80	5.92
CO	2.74	0.90	0.56	7.44	3.30	1.00	0.15	2.43	4.00	1.00	0.02	2.00	2.88	0.84	0.65	6.47
EC	–	–	–	–	2.32	1.00	0.06	17.00	–	–	–	–	2.32	0.50	0.06	17.00
MX	2.96	0.88	0.25	3.69	3.14	1.00	0.02	1.43	3.72	0.67	0.01	2.00	3.00	0.87	0.27	3.51
SV	3.22	1.00	0.03	1.00	–	–	–	–	–	–	–	–	3.22	1.00	0.03	1.00
PY	3.38	1.00	0.16	9.11	3.84	1.00	0.04	11.40	3.96	1.00	0.01	2.00	3.60	0.96	0.20	9.35
UY	3.24	1.00	0.29	2.40	–	–	–	–	–	–	–	–	3.24	1.00	0.29	3.24
VE	3.80	1.00	0.36	3.27	3.68	0.97	0.33	2.39	–	–	–	–	3.64	0.99	0.69	2.88
ALL	3.34	0.69	0.35	4.57	3.62	0.97	0.15	2.82	3.66	0.74	0.03	2.44	3.36	0.73	0.51	4.08

Table 2. Country Breakdown of Forecasting Performance for Different Data Sources.

QS = Quality Score; Pr = Precision; Rec = Recall; LT = Lead Time. AR = Argentina; BR = Brazil; CL = Chile; CO = Colombia; EC = Ecuador; SV = El Salvador; MX = Mexico; PY = Paraguay; UY = Uruguay; VE = Venezuela. A – indicates that the source did not produce any warnings for that country in the studied period.

event's reported location are within 300 kilometers of each other (the distance offset), and (3) the forecast event date is within a given interval of the true event date (the date offset). Once these inclusion criteria apply, the quality score (QS) of the match is defined as a combination of the location score (LS) and date score (DS):

$$QS = (LS + DS) * 2 \quad (3)$$

where

$$LS = 1 - \frac{\min(distanceoffset, 300)}{300} \quad (4)$$

and

$$DS = 1 - \frac{\min(dateoffset, INTERVAL)}{INTERVAL} \quad (5)$$

Here, we explore INTERVAL values from 0 to 7. If an alert (conversely, GSR event) cannot be matched to any GSR event (alert, respectively), these

Source		AR	BR	CL	CO	EC	SV	MX	PY	UY	VE	All
News/Blogs	LS	0.82	0.76	0.75	0.60	–	0.75	0.66	0.79	0.79	0.95	0.81
	DS	0.75	0.81	0.78	0.77	–	0.86	0.82	0.90	0.83	0.95	0.86
Facebook	LS	1.0	0.92	–	1.00	–	–	0.86	0.98	–	–	0.93
	DS	0.85	0.89	–	1.00	–	–	1.00	1.00	–	–	0.90
Twitter	LS	0.88	–	0.84	0.81	0.45	–	0.71	0.98	–	0.91	0.89
	DS	0.88	–	0.92	0.84	0.71	–	0.86	0.94	–	0.93	0.92

Table 3. Comparing the Location and Date Scores of Different Sources in Specific Countries.

AR = Argentina; BR = Brazil; CL = Chile; CO = Colombia; EC = Ecuador; SV = El Salvador; MX = Mexico; PY = Paraguay; UY = Uruguay; VE = Venezuela. A – indicates that the source did not produce any warnings for that country in the studied period.

unmatched alerts (and events) will negatively affect the precision (and recall) of the system. The lead time, for a matched alert-event pair, is calculated as the difference between the date on which the forecast was made and the date on which the event was reported (this should not be confused with the date score, which is the difference between the predicted event date and the actual event date). Lead time concerns itself with reporting and forecasting, whereas the date score is concerned with quality or accuracy. We conduct a series of experiments to evaluate the performance of our system.

How does the distribution of protests detected by the system compare with the actual distribution of protests in the GSR? Figure 5 reveals pie charts of both distributions. As shown, Mexico, Brazil, and Venezuela experience the lion's share of protests in our region of interest, and the protests detected also match these modes although not the specific percentages. Smaller countries like Ecuador, El Salvador, and Uruguay do experience protests but they are not as prominently detected as those for other countries; we attribute this to their smaller social media footprint (relative to countries like Brazil and Venezuela).

Are there country-specific selective superiorities for the different data sources considered here? Table 2 presents a breakdown of performance, countrywise and sourcewise, of our approach for a particular month (March 2014). It is clear that the multiple data sources are necessary to achieve a high recall and that by and large these sources are providing mutually exclusive alerts (note also that some data sources do not produce alerts for specific countries). Between Twitter and Facebook, the former is a better source of alerts for countries like Chile and the latter is a better source for Argentina, Brazil, Colombia, and Mexico. News and blogs achieve higher recall than social media sources indicating that most plans for protests are announced in established media. They are also higher quality sources for alerts in countries like El Salvador, Paraguay, and Uruguay. Finally, note that news and blogs offer a much higher lead time (4.57 days) as compared to that for Facebook (2.44 days) or for Twitter (2.82 days). The quality scores are further broken down in table 3 into their date and location components.

A longitudinal perspective on quality scores is given in figure 6a. Note that, in general, Twitter tends to have a higher quality score as multiple retweets of future event mentions is a direct indicator of the popularity of an event as well as the intent of people to join an event. In contrast, mentions of future events in news do not directly shed any insight into popularity or people's support for the event's causes.

How did our system fare in detecting key countrywide protests? The recent Venezuelan protests against President Nicolas Maduro and the Brazilian protests during June 2013 against bus-fare hikes were two significant protests during our period of evaluation. Figures 6b and 6f describe our performance under these two situations illustrating the count of protests detected against the GSR. Notice that our system was able to identify the Venezuelan protests much better than the Brazilian protests. This is because there was a significant amount of spontaneity to the Brazilian protests; they arose as discontent about bus-fare increases but later morphed into a broader set of protests against government and most of these subsequent protests were not planned.

What is the trade-off between lead time and quality? Figure 6c shows that the QS of the planned protest model decreases (as expected) with lead time, initially, but later rises again. The higher quality scores toward the right of figure 6c are primarily due to Facebook event pages.

How does the method perform under stringent matching criteria? Figure 6d shows the performance of the model when the matching window is varied from 7 to 1 in steps. We can see that the performance

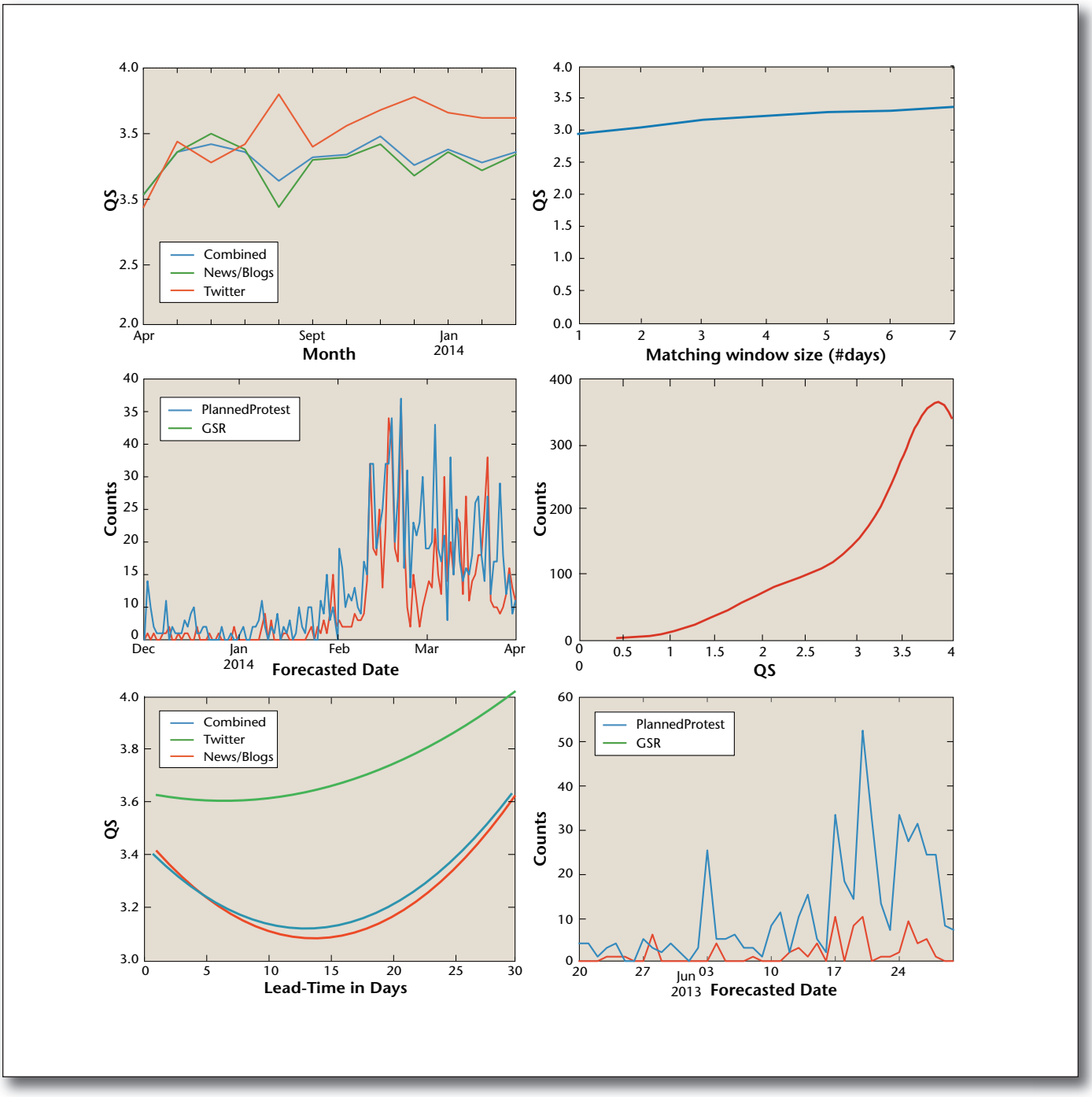

Figure 6. Evaluation of Planned Protest Forecasting System.

degrades quite gracefully even under the strict matching interval of a one-day difference.

What is the distribution of quality scores? The clear mode toward the right side of figure 6e signifies that a majority of the planned protest alerts are of high quality. Further, the quality score distribution is unimodal suggesting that the careful reasoning of locations and date normalization is crucial to achieving high quality.

Development and Maintenance

The core algorithms behind the planned protest detector were implemented in Python. The PSL geocoder was implemented in Java. Among the external libraries utilized, the Basis Rosette Linguistic Platform is the key library. The development process took three months (June 2012 to August 2012) and was primarily led by the first author with contribu-

tions from the other authors. After two months of testing (September 2012 and October 2012), the system was deployed in November 2012 on the commercial Amazon Web Services cloud infrastructure in a cluster configuration. More details about the EMBERS system architecture can be found in the paper by Doyle et al. (2014). Alerts generated by the system are automatically emailed to Mitre (for evaluation) as well as to analysts who are subject matter experts in Latin America.

Because there is no explicit training phase, the system has required minimal reengineering over time. Key changes made to the system over time were to increase the sources used for data ingestion and supporting the inclusion of additional phrases. Agile software engineering methods were used for project management. We estimate the effort to maintain the system as 0.25 persons (software engineer) per year, which entails keeping data sources current, ensuring that the phrase list continues to stay relevant, and performing periodic checks and evaluations of the sytem.

Discussion

We have described an approach to forecasting protests by detecting mentions of future events in news and social media. The twin issues of resolving the date and resolving the location have been addressed satisfactorily to realize an effective protest forecasting system. As different forms of communication media gain usage, systems like ours will be crucial to understanding the concerns of citizens.

Our future work is aimed at three areas. First, to address situations such as nationwide protests and systems of protests, we must generalize our system from generating protests at a single article level to digesting groups of articles. Doing so would require more sophisticated probabilistic reasoning, which we believe can be done using PSL. Second, we would like to generalize our approach, which currently does detection of overt plans for protest, to not-so-explicitly stated expressions of discontent. Finally, we plan to consider other population-level events of interest than just civil unrest, for example, domestic political crises, and design detectors to recognize the imminence of such events.

Acknowledgments

Research for this article has been supported by the Intelligence Advanced Research Projects Activity (IARPA) through DoI/NBC contract D12PC000337. The U.S. government is authorized to reproduce and distribute reprints of this work 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 U.S. government.

Notes

1. See www.basistech.com/text-analytics/rosette.

References

Bach, S. H.; Huang, B.; London, B.; and Getoor, L. 2013. Hinge-Loss Markov Random Fields: Convex Inference for Structured Prediction. In *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Ninth Conference*. Corvallis, OR: AUAI Press.

Baeza-Yates, R. 2005. Searching the Future. Paper presented at the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval, Salvador, Brazil August 15–19.

Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; and Etzioni, O. 2007. Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conferences on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Chambers, N., and Jurafsky, D. 2011. Template-Based Information Extraction Without the Templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT*. Stroudsburg, PA: Association for Computational Linguistics.

Compton, R.; Lee, C.; Lu, T.-C.; De Silva, L.; and Macy, M. 2013. Detecting Future Social Unrest in Unprocessed Twitter Data: "Emerging Phenomena and Big Data." In *Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics, ISI*. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Doyle, A.; Katz, G.; Summers, K. M.; Ackermann, C.; Zavorin, I.; Lim, Z.; Muthiah, S.; Zhao, L.; Lu, C.; Butler, P.; Khandpur, R. P.; Fayed, Y.; and Ramakrishnan, N. 2014. The EMBERS Architecture for Streaming Predictive Analytics. In *Proceedings of the IEEE 2014 International Conference on Big Data*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/BigData.2014.7004477

Gabrilovich, E.; Dumais, S.; and Horvitz, E. 2004. Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. In *Proceedings of the 13th International Conference on World Wide Web (WWW)*. New York: Association for Computing Machinery.

Hurriyetoglu, A. H.; Kunneman, F.; and van den Bosch, A. 2013. Estimating the Time Between Twitter Messages and Future Events. In *Proceedings of the Dutch-Belgian Workshop on Information Retrieval,* CEUR Workshop Proceedings Volume 986. Aachen, Germany: RWTH Aachen University.

Jatowt, A., and Au Yeung, C. M. 2011. Extracting Collective Expectations About the Future from Large Text Collections. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*. New York: Association for Computing Machinery. dx.doi.org/10.1145/2063576.2063759

Kallus, N. 2014. Predicting Crowd Behavior with Big Public Data. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*. New York: Association for Computing Machinery.

Kawai, H.; Jatowt, A.; Tanaka, K.; Kunieda, K.; and Yamada, K. 2010. ChronoSeeker: Search Engine for Future and Past Events. In *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication, ICUIMC*. New York: Association for Computing Machinery.

Kimmig, A.; Bach, S.; Broecheler, M.; Huang, B.; and Getoor, L. 2012. A Short Introduction to Probabilistic Soft Logic.

Paper presented at the NIPS Workshop on Probabilistic Programming: Foundations and Applications, South Lake Tahoe, CA, Dec. 7–8.

Llorens, H.; Derczynski, L.; Gaizauskas, R. J.; and Saquete, E. 2012. TIMEN: An Open Temporal Expression Normalisation Resource. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012*). Paris: European Language Resources Association.

Mani, I., and Wilson, G. 2000. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics,* ACL. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/1075218.1075228

Padró, L., and Stanilovsky, E. 2012. Freeling 3.0: Towards Wider Multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Paris: European Language Resources Association.

Pustejovsky, J.; Castano, J. M.; Ingria, R.; Sauri, R.; Gaizauskas, R. J.; Setzer, A.; Katz, G.; and Radev, D. R. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *In New Directions in Question Answering,* ed. M. Maybury. Menlo Park, CA: AAAI Press.

Radinsky, K., and Horvitz, E. 2013. Mining the Web to Predict Future Events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM*. New York: Association for Computing Machinery. dx.doi.org/10.1145/2433396.2433431

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, C.; 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. 2014. "Beating the News" with EMBERS: Forecasting Civil Unrest Using Open Source Indicators. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery. dx.doi.org/10.1145/2623330.2623373

Riloff, E., and Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/1119355.1119369

Strötgen, J.; Armiti, A.; Van Canh, T.; Zell, J.; and Gertz, M. 2014. Time for More Languages: Temporal Tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing* 13(1):1–21. dx.doi.org/10.1145/2540989

Tops, H.; van den Bosch, A.; and Kunneman, F. 2013. Predicting Time-to-Event from Twitter Messages. Paper presented at the 2013 Benelux Conference on Artificial Intelligence, BNAIC, Delft, The Netherlands, Nov. 7–8.

Truvé, S. 2011. Big Data for the Future: Unlocking the Predictive Power of the Web. White Paper, Recorded Future, Inc., Cambridge, MA.

Verhagen, M.; Gaizauskas, R.; Schilder, F.; Hepple, M.; Moszkowicz, J.; and Pustejovsky, J. 2009. The TempEval Challenge: Identifying Temporal Relations in Text. *Language Resources and Evaluation* 43(2): 161–179. dx.doi.org/10.1007/s10579-009-9086-z

Xu, J.; Lu, T.-C.; Compton, R.; and Allen, D. 2014. Civil Unrest Prediction: A Tumblr-Based Exploration. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference*. Berlin: Springer. dx.doi.org/10.1007/978-3-319-05579-4_49

Sathappan Muthiah is a Ph.D. student in computer science at the Virginia Polytechnic Institute and State University. He is a research assistant at the Discovery Analytics Center. His research interests include spatial and temporal event detection, text mining, and information retrieval. He received his B.Tech in information technology from the National Institute of Technology, Bhopal, India.

Bert Huang is an assistant professor in the Department of Computer Science at the Virginia Polytechnic Institute and State University. Huang's research investigates machine learning, with a focus on topics including structured prediction, statistical relational learning, and computational social science. He is an action editor for the Journal of Machine Learning Research, and his papers have been published at conferences including NIPS, ICML, UAI, and AISTATS. He earned his Ph.D. from Columbia University in 2011, and he was a postdoctoral research associate at the University of Maryland, College Park.

Jaime Arredondo is a Ph.D. candidate in global public health at the University of California, San Diego. His research centers on developing analytical methods and techniques to analyze the effect of policing practices as a major driver of HIV transmission, particularly in the U.S.–Mexico border region.

David Mares is a professor of political science at the University of California, San Diego, where he holds the Institute of the Americas Chair for Inter-American Affairs. He is the author or editor of 10 books and his publications have appeared in English, Spanish, French, Portuguese, Italian, and Chinese. His research and teaching interests include international conflict, Latin American energy politics, the political economy of drug policy, and civil-military relations.

Lise Getoor is a professor in the Computer Science Department at the University of California, Santa Cruz. Her research areas include machine learning, data integration, and reasoning under uncertainty, with an emphasis on graph and network data. She is a Fellow of the Association for Artificial Intelligence, recipient of nine best paper and best student paper awards, and was cochair of ICML 2011. She received her Ph.D. from Stanford University in 2001, her MS from the University of California, Berkeley, and her BS from the University of California, Santa Barbara, and was a professor in the Computer Science Department at the University of Maryland, College Park from 2001–2013.

Graham Katz earned his Ph.D. in linguistics and computational linguistics from the University of Rochester. He was an applied research manager at CACI Inc. from 2012 to 2015 where he led the implementation of text-enrichment pipelines for EMBERS. Katz is now with IBM Inc.

Naren Ramakrishnan is the Thomas L. Phillips Professor of Engineering at the Virginia Polytechnic Institute and State University. He directs the Discovery Analytics Center, a university center pursuing advanced research in data mining and knowledge discovery with applications to important areas of national interest. He is also the principal investigator of the ongoing IARPA Open Source Indicators project aimed at forecasting significant societal events (disease outbreaks, civil unrest, elections) from open source data sets. He received his Ph.D. in computer sciences from Purdue University.